63-1-5

MEMORANDUM
RM-3866-PR
NOVEMBER 1962

DERIVATION OF ESTIMATING RELATIONSHIPS: AN ILLUSTRATIVE EXAMPLE

G. H. Fisher

PREPARED FOR:
UNITED STATES AIR FORCE PROJECT RAND

The RAND Corporation
SANTA MONICA • CALIFORNIA

MEMORANDUM
RM-3366-PR
NOVEMBER 1962

DERIVATION OF ESTIMATING RELATIONSHIPS: AN ILLUSTRATIVE EXAMPLE

G. H. Fisher

This research is sponsored by the United States Air Force under Project RAND — Contract No. AF 49(638)-700 — monitored by the Directorate of Development Planning, Deputy Chief of Staff, Research and Technology, Hq USAF. Views or conclusions contained in this Memorandum should not be interpreted as representing the official opinion or policy of the United States Air Force. Permission to quote from or reproduce portions of this Memorandum must be obtained from The RAND Corporation.

The RAND Corporation
1700 MAIN ST • SANTA MONICA • CALIFORNIA

PREFACE

The RAND Corporation has been assisting the Air Force Systems Command in developing and teaching a course in military systems cost analysis concepts and techniques given at the Air Force Institute of Technology. This Memorandum was prepared for that course and is being published as one of a series of memorandums which will serve as course material for future classes. The paper should also be of interest to others within the Air Force who are concerned with the problem of derivation of estimating relationships.

Basic data used in the statistical analyses presented in this Memorandum were taken from actual historical data sources. However, the data have been transformed and adjusted to eliminate security and proprietary information classifications and to better serve the requirements for use as instructional material. For this reason the basic data should be regarded as essentially hypothetical, and the estimating relationships derived therefrom must be viewed as illustrative only. They should not be used in actually making estimates of airframe initial tooling cost.

SUMMARY

This Memorandum presents illustrative examples of how statistical regression analysis may be used to derive estimating relationships from historical data. The specific illustration pertains to estimating relationships for airframe initial tooling cost as a function of aircraft performance and physical characteristics.

Examples of simple linear regression, logarithmic linear regression, second degree regression, and multiple linear regression analyses are presented and discussed.

Student problems are contained in Appendices B, C, and D.

CONTENTS

I. INTRODUCTION

The purpose of this Memorandum is to present a step-by-step illustration of how statistical regression analysis may be used in the derivation of estimating relationships from historical data. It is to be used as part of the instructional material for the USAFIT cost analysis course and possibly for cost analysis training that might be given at RAND.

In this case the specific example refers to an estimating relationship for airframe initial (non-recurring) tooling cost for manned aircraft. The objective is to show how initial tooling cost may be related to aircraft characteristics. Information and data used as a basis for the statistical analyses were taken from actual historical data sources. However, the data have been transformed and adjusted to eliminate security and proprietary information classifications and to better serve the requirements for use as instructional material. For this reason the basic data should be regarded as essentially hypothetical, and the estimating relationships derived therefrom must be viewed as illustrative only. They should not be used in actually making estimates of airframe initial tooling cost. The objective here is solely to demonstrate manipulation of data and analytical techniques.

Even though the information underlying the analysis must be considered hypothetical, the example does contain and illustrate many of the major problems encountered in a real life situation -- for example, a very small sample size, uneven distribution of observations in the sample over the ranges of the explanatory variables, the difficulty of ascertaining reasonably good explanatory variables, non-homogeneities

in the data, and the like. All of these factors, and others, compound the problems involved in derivation of statistical estimating relationships for use in military cost analysis activities. While this is also true in other fields, the difficulties seem particularly severe in the cost analysis of advanced weapon systems and forces. But we must do the best we can with a small and very often heterogeneous data base.

It should also be remembered that statistical estimating relationships derived in such an adverse environment must always be used with caution, particularly when extrapolating to distant future weapon systems. For the most part, use of an estimating relationship should be viewed as a point of departure, to be modified by experience, judgment, and external or supplementary information. This in no way downgrades the need for developing and keeping current a good library of estimating relationships. Without such a library, the cost analyst does not have even a point of departure. Also, a reasonably complete stock of estimating relationships is a prime prerequisite to being able to do sensitivity analysis studies.

II. STATEMENT OF THE PROBLEM

Suppose that we have collected historical data on airframe initial tooling cost for 14 types of aircraft: 7 fighters (F-1, F-2, ..., F-7) and 7 bombers (B-1, B-2, ..., B-7). (To eliminate the effect of price level changes, these cost data were adjusted statistically and expressed in terms of 1962 dollars.) In addition certain aircraft characteristics data have been assembled for each of the 14 cases: AMPR airframe weight, maximum speed, and combat radius. All of this information is summarized in Table 1 on the next page.

Given the data contained in Table 1, the problem is to try to derive an estimating relationship for initial tooling cost (X_1) expressed as a function of one or more of the "explanatory" variables X_2, X_3, X_4. We are immediately confronted with questions like the following: What explanatory variable or variables shall we include in the estimating relationships? What functional form seems appropriate? Shall we stratify the sample -- e.g., treat bombers and fighters separately?

Several techniques are available to help answer these questions. Probably the simplest way to proceed is merely to plot the basic data on scatter diagrams -- a separate plot for each of the explanatory variables (X_2, X_3, X_4) in relation to the dependent variable (X_1). This has been done in Figs. 1 - 3.

A brief examination of Figs. 1 - 3 suggests that:

(1) Of the three explanatory variables, airframe weight (X_2) seems to best "explain" variations in initial tooling cost (X_1), combat radius (X_4) is second best, and maximum speed is the least satisfactory.

Table 1

INITIAL TOOLING COST AND VARIOUS AIRCRAFT CHARACTERISTIC DATA

Aircraft Type	Initial Tooling Cost (Millions 1962 $) ($X_1$)	AMPR Airframe Weight (M of Lbs) (X_2)	Maximum Speed (Kn) (X_3)	Combat Radius (N Mi) (X_4)
F-1	8	7	525	275
F-2	15	8	575	300
F-3	20	9	600	250
F-4	40	15	750	300
F-5	30	12	800	600
F-6	35	20	1100	360
F-7	70	25	1200	550
B-1	50	40	525	1800
B-2	265	115	550	3000
B-3	110	50	1100	2200
B-4	85	70	525	1000
B-5	60	50	330	1000
B-6	20	20	500	800
B-7	165	90	550	780

SOURCE: Hypothetical data.

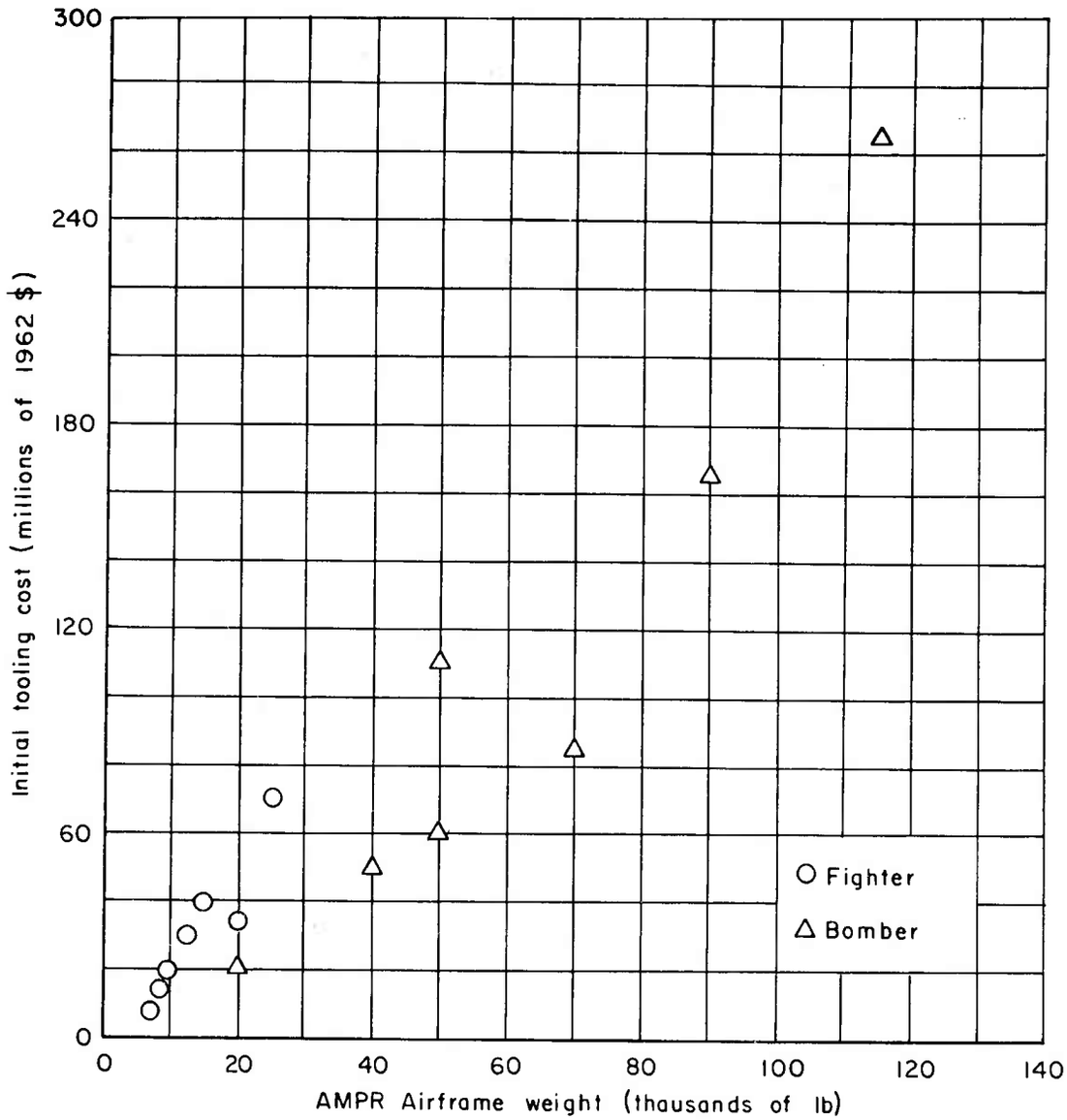

Fig. 1— Initial tooling cost versus airframe weight

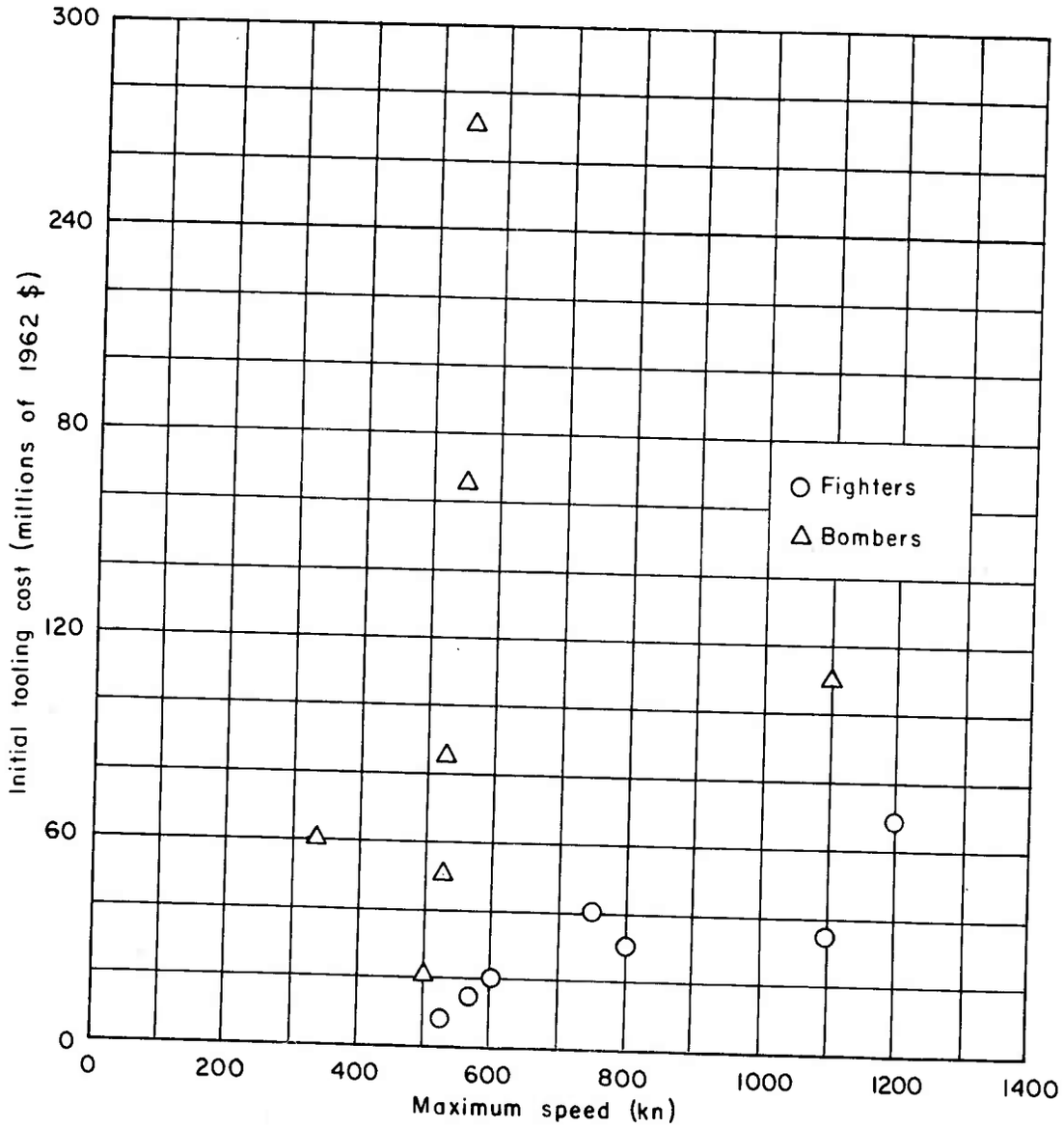

Fig. 2 — Initial tooling cost versus maximum speed

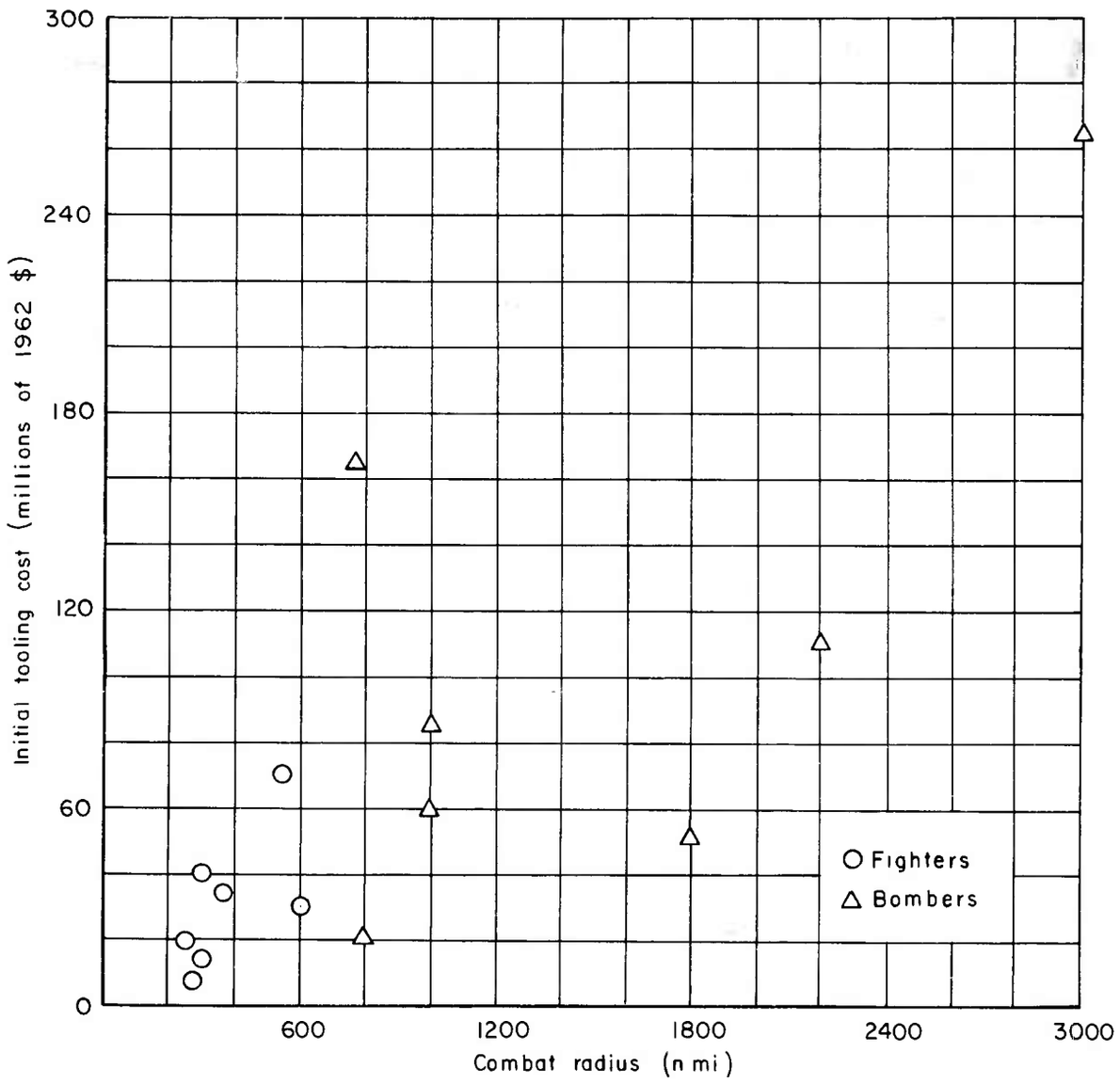

Fig. 3 — Initial tooling cost versus combat range

(2) In the case of X_1 vs. X_2 and X_1 vs. X_4, a linear functional form seems about as appropriate as anything else.

(3) Except possibly for Fig. 2 (X_1 vs. X_3), there does not seem to be a compelling reason for stratification -- e.g., treating the bombers and fighters separately. In Figs. 1 and 3, bombers and fighters tend to be in the same general path of an estimating relationship that would be fitted to the data. Also in the specific case at hand, the total sample (14 observations) is already small, and to divide the sample into bombers and fighters would reduce the data to two sub-sets of only 7 observations each. We shall therefore treat bombers and fighters together in the statistical analyses which follow.

(4) In general, the scatter diagrams do not indicate as close a relationship between X_1 and the explanatory variables as we would like. But this is rather typical; and we have to proceed to develop the best estimating relationships we can, given the available data. Also, in using relationships so derived, care must be exercised to take into account the limitations of the data base.

III. LINEAR NORMAL REGRESSION ANALYSIS OF INITIAL TOOLING COST AS A FUNCTION OF AIRFRAME WEIGHT

As an illustrative example, we shall now proceed to derive an estimating relationship for initial tooling cost as a function of airframe weight, using the data contained in the first two columns of Table 1. The specific statistical technique used will be a linear normal regression model.*

The first step is to take the basic data for X_1 and X_2 from Table 1 and compute the cross products and squares, the sums of these items, and the sample means for X_1 and X_2. The results of these calculations are shown in Table 2 on the next page.

These data are now used to compute estimates of the parameters α and β in the linear regression (estimating) equation:

$$(1) \qquad X_1 = \alpha + \beta X_2$$

In a linear normal regression model this amounts to finding the values of α and β such that the sum of the squares of the deviations of the sample observations from the regression line will be at a minimum; i.e.,

$$(2) \qquad \Sigma\left[X_1 - (\alpha + \beta X_2)\right]^2 = \text{minimum}$$

The minimization of (2) with respect to α and β is a straightforward calculus problem. The results of such a minimization yield the

---

*For a detailed discussion of linear normal regression models, see A. M. Mood, Introduction to the Theory of Statistics, McGraw-Hill Book Company, Inc., 1950, pp. 291-299; F. E. Croxton and D. J. Cowden, Applied General Statistics, Prentice-Hall, Inc., 1940, Chap. XXII; and G. W. Snedecor, Statistical Methods, Iowa State College Press, Fourth Edition, pp. 103-137.

Table 2

DATA FOR REGRESSION ANALYSIS OF INITIAL TOOLING COST AND AIRFRAME WEIGHT

Aircraft Type	X_1 (Tooling Cost)	X_2 (Airframe Weight)	X_1X_2	X_1^2	X_2^2
F-1	8	7	56	64	49
F-2	15	8	120	225	64
F-3	20	9	180	400	81
F-4	40	15	600	1,600	225
F-5	30	12	360	900	144
F-6	35	20	700	1,225	400
F-7	70	25	1,750	4,900	625
B-1	50	40	2,000	2,500	1,600
B-2	265	115	30,475	70,225	13,225
B-3	110	50	5,500	12,100	2,500
B-4	85	70	5,950	7,225	4,900
B-5	60	50	3,000	3,600	2,500
B-6	20	20	400	400	400
B-7	165	90	14,850	27,225	8,100
Totals	973	531	65,941	132,589	34,813

SOURCE: Table 1.

Sample size = N = 14

$$\text{Mean of the } X_1\text{'s} = \frac{\Sigma X_1}{N} = \frac{973}{14} = 69.50$$

$$\text{Mean of the } X_2\text{'s} = \frac{\Sigma X_2}{N} = \frac{531}{14} = 37.93$$

so-called "normal equations" for linear normal regression:

$$(3) \qquad \Sigma X_1 = N\alpha + \beta \Sigma X_2$$

$$(4) \qquad \Sigma X_1 X_2 = \alpha \Sigma X_2 + \beta \Sigma X_2^{2*}$$

The relevant numerical values to be substituted in (3) and (4) are contained in Table 2. They are:

$$N = 14$$
$$\Sigma X_1 = 973$$
$$\Sigma X_2 = 531$$
$$\Sigma X_1 X_2 = 65,941$$
$$\Sigma X_2^2 = 34,813$$

Substituting these numbers in the normal equations (3) and (4), we obtain:

$$(5) \qquad 973 = 14\alpha + 531\beta$$

$$(6) \qquad 65,941 = 531\alpha + 34,813\beta$$

Equations (5) and (6) must be solved simultaneously to obtain estimates of the regression coefficients (α and β). To do this we must first eliminate one of the variables α or β. Let us arbitrarily select α for elimination. First, calculate the ratio of the coefficients of α in equations (5) and (6):

---

*A derivation of the normal equations is contained in Appendix A.

$$\frac{531}{14} = 37.928571.$$

Then multiply equation (5) by 37.928571, obtaining a new equation (5'):

$$(37.928571)(973) = (37.928571)(14\alpha + 531\beta),$$

or,

$$(5') \qquad 36{,}904.4996 = 531\alpha + 20{,}140.0712\beta.^{*}$$

Finally, by subtracting equation (5') from equation (6) we can eliminate α and solve the resulting equation for β:

$$\begin{array}{ll} (6) & 65{,}941.0000 = 531\alpha + 34{,}813.0000\,\beta \\ (5') & \underline{-36{,}904.4996 = -531\alpha - 20{,}140.0712\,\beta} \\ & 29{,}036.5004 = \qquad 14{,}672.9288\,\beta \end{array}$$

$$\beta = \frac{29{,}036.5004}{14{,}672.9288} = 1.978916.$$

Having an estimate of the regression coefficient β, the estimate of α may be calculated by substituting $\beta = 1.978916$ in equation (6) and solving the resulting equation for α:

$$\begin{aligned} 65{,}941 &= 531\alpha + (34{,}813)(1.978916) \\ 65{,}941 &= 531\alpha + 68{,}892.0027 \\ 531\alpha &= -2951.0027 \\ \alpha &= -5.557444. \end{aligned}$$

The results of these calculations may be checked by substituting $\alpha =$

---

*Notice that the value of equation (5) is unchanged, since we multiplied both sides of the equation by the same number (37.928571).

-5.557444 and $\beta = 1.978916$ in equation (5):

$$973 = (14)(-5.557444) + (531)(1.978916)$$

$$973 = -77.804216 + 1{,}050.804396$$

$$973 \cong 973.000180$$

Thus, the calculated regression coefficients are:

$$\alpha = -5.557444$$

$$\beta = 1.978916,$$

and the estimating equation is:

$$(7) \qquad X_1 = -5.5574 + 1.9789X_2,$$

where

X_1 = Initial tooling cost in millions of 1962 dollars,

X_2 = AMPR airframe weight in thousands of pounds.

Equation (7) may be plotted on the scatter diagram contained in Fig. 1. Two plot points are needed for this purpose. Computing the value of X_1 for $X_2 = 10$ and $X_2 = 100$ from equation (7), we obtain:

$$X_1 = 14.2 \text{ (for } X_2 = 10)$$

$$X_1 = 192.3 \text{ (for } X_2 = 100)$$

The results of plotting these numbers and drawing in the regression

line is shown in Fig. 4 (the solid line).*

The regression line is in effect an <u>average</u> relationship. Specifically in this instance it is that line about which the sum of the squares of the deviations of X_1 is at a minimum. Usually, however, we are not only interested in averages, but also in the <u>reliability</u> of these averages. In the case of regression analysis, one measure of reliability is the <u>standard error of estimate</u> (S) of the regression equation. The standard error of estimate is defined as the square root of the <u>unexplained</u> variance of the X_1's in the sample. This unexplained variance is obtained by computing the difference between the total variance of the X_1's and the "explained" variance (the variance accounted for by the regression line).** The shortcut method for determining the unexplained variance is:

$$(8) \qquad s^2 = \frac{\Sigma X_1^2 - (\alpha \Sigma X_1 + \beta \Sigma X_1 X_2)}{N} \text{***}$$

The <u>unadjusted</u> standard error of estimate (S) is the square root of expression (8). The <u>adjusted</u> standard error of estimate ($\bar{S}$) is obtained by subtracting the number of parameters in the regression

---

*The fact that the "least squares" fit to the sample data produces a slightly negative value for α may seem disturbing. This need not be the case, however, since we would not want to use the estimating equation for extremely low values of X_2 -- certainly not for $X_2 < 5000$ lb. Also, the standard error of α is such that in a statistical inference sense we cannot be confident that the <u>universe</u> value of α is less than zero.

**The concepts of total, explained, and unexplained variance are discussed in more detail later.

***See F. E. Croxton and D. J. Cowden, <u>Applied General Statistics</u>, Prentice-Hall, Inc., 1940; pp. 661-63 and 671.

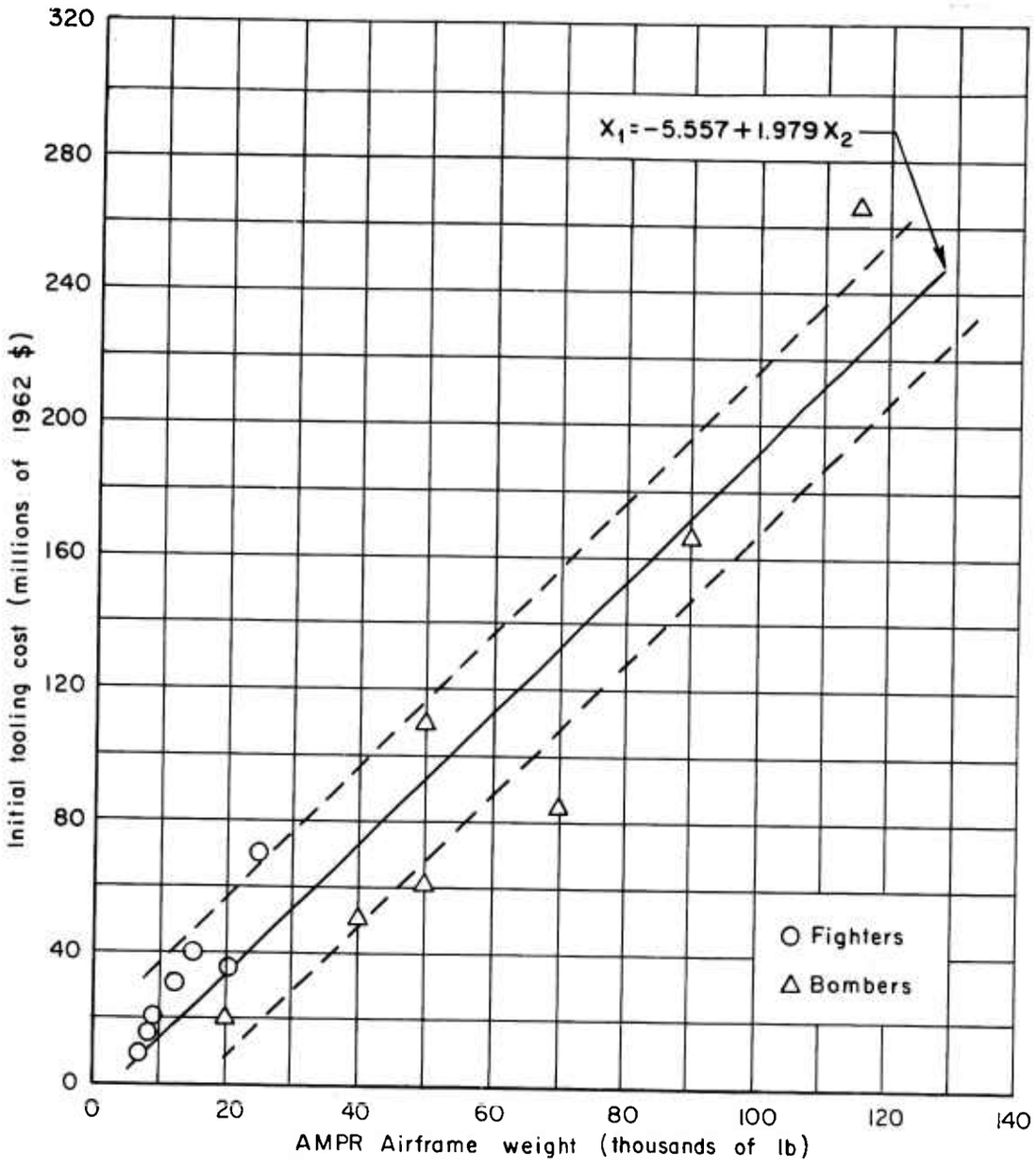

Fig. 4 — Initial tooling cost versus airframe weight

equation from the sample size (N) in the formula for S. In the case of simple normal linear regression, the number of parameters in the regression equation is 2. Therefore, the formula for $\bar{S}$ is:

$$(9) \qquad \bar{S} = \sqrt{\frac{\Sigma X_1^2 - (\alpha \Sigma X_1 + \beta \Sigma X_1 X_2)}{N - 2}}$$

From (9) it is clear that for large sample sizes (large N) the adjustment is of no importance. However, in small samples -- particularly very small samples -- the adjustment can make quite a difference. In general $S < \bar{S}$, and S approaches $\bar{S}$ as N becomes large.

Regarding interpretation of the standard error of estimate, the main point is that in normal linear regression analyses one might expect that about two-thirds of the sample observations would fall within a region bounded by $\pm$ 1 standard error of estimate from the regression line; about 95 per cent of the observations within $\pm$ 2 standard error of estimate from the regression line; and virtually all of the observations within $\pm$ 3 standard error of estimate. In practice these generalizations do not tend to hold up very well in very small sample cases. It should also be emphasized that here we are talking about distribution of the observations <u>in the sample</u>; and not about the reliability or "confidence limits" pertaining to a predicted X_1 as given by the estimating equation. (The subject of prediction intervals will be taken up later.)

Returning now to our illustrative example, the standard error of estimate is computed by substituting the appropriate data in equation (9). We have already calculated $\alpha = -5.557444$ and $\beta = 1.978916$; the sample size is 14; and the required summations are contained in

Table 2. We have, therefore:

$$\bar{S} = \sqrt{\frac{132{,}589 - (-5.557444)(973) - (1.978916)(65{,}941)}{14 - 2}}$$

$$= \sqrt{\frac{132{,}589 + 5{,}407.39 - 130{,}491.70}{12}}$$

$$= \sqrt{\frac{7{,}504.69}{12}} = \sqrt{625.39}$$

$$= 25.01 \text{ (millions of 1962 dollars).}^{*}$$

In Fig. 4, a band of $\pm 1\ \bar{S}$ from the regression line has been plotted on the scatter diagram (the dotted lines).

For some purposes -- particularly in comparing one $\bar{S}$ with another -- it is useful to compute a relative standard error of estimate. One such measure is the coefficient of variation (C), which relates the standard error of estimate to the mean of the sample X_1's:

$$(10) \qquad C = \frac{\bar{S}}{\bar{X}_1}.$$

In the case of our illustrative example, the mean of the X_1's (from Table 2) is $\bar{X}_1 = 69.5$. The value of C, therefore, is:

$$C = \frac{25.01}{69.50} = 36\%.$$

The question of reliability of an estimating equation is, of course,

*As a matter of interest, the unadjusted standard error of estimate is 23.16.

a relative matter -- i.e., relative to the context in which the equation is to be used. However, as a general rule, we would usually prefer not to have a C as high as 36 per cent. Something like 10 to 20 per cent would be more desirable.

So far the question of reliability has been considered in the context of the regression equation in relation to the sample observations. But this is usually not the context that is of greatest interest. Rather than being concerned with how well the regression equation describes the sample observations per se, the analyst is most usually interested in using the estimating equation to predict values of X_1 in the "population" or "universe" that the sample supposedly represents. In the context of prediction, the standard error of estimate does not furnish a good measure of uncertainty or reliability of the estimating (regression) equation. In a formal sense, what we would like is somewhat as follows. For a given value of the explanatory variable, say $\hat{X}_2$, the estimating equation is used to obtain a predicted value of the dependent variable:

$$\hat{X}_1 = \alpha + \beta \hat{X}_2 .$$

Then we would like to put a boundary around $\hat{X}_1$ -- say $\hat{X}_1 \pm A$ -- such that there is a certain level of confidence that the established interval does indeed "bracket" the "true" value of X_1 in the population. The subject of "prediction intervals" is addressed to this problem.

In the case of normal linear regression it has been established that a $100(1 - \epsilon)$ per cent prediction interval for an estimated value

of the dependent variable, say $\hat{X}_1$, can be constructed as follows:*

$$\hat{X}_1 \pm A,$$

where

$$(11) \qquad A = \bar{S} t_\epsilon \sqrt{\frac{N}{N-2}\left[\frac{N+1}{N} + \frac{(\hat{X}_2 - \bar{X}_2)^2}{\Sigma (X_2 - \bar{X}_2)^2}\right]}$$

The meaning of the notation in (11) is:

$\bar{S}$ = standard error of the estimating equation from which $\hat{X}_1$ was obtained,

t_ϵ = the value of t obtained from a table of Student's "t" distribution for the ϵ significance level,

N = size of the sample used to derive the estimating equation,

$\hat{X}_2$ = the specified value of the explanatory variable used as a basis for obtaining $\hat{X}_1$,

$\bar{X}_2$ = the mean of the X_2's in the sample used to derive the estimating equation,

$\Sigma (X_2 - \bar{X}_2)^2$ = the sum of squared deviations of the sample X_2's from their mean.

We shall now apply this procedure to our illustrative example, using the estimating equation (7) derived previously:

$$(7) \qquad X_1 = -5.5574 + 1.9789X_2,$$

and assuming that we want to estimate the value of X_1 for X_2 = 90,000 lb. From (7) we have:

---

*For a derivation and explanation of this procedure, see A. M. Mood, op. cit., pp. 297-99.

$$\hat{X}_1 = -5.5574 + 1.9789(90),$$

$$= -5.5574 + 178.1010,$$

$$= 172.5 \text{ (millions of 1962 dollars).}$$

Now let us assume that we want to establish a 95 per cent prediction interval around $\hat{X}_1 = 172.5$, using equation (11) to derive the value of A. The necessary data are as follows:

$\overline{S} = 25.01$. This is the value of the standard error of estimate calculated previously for equation (7).

$\epsilon = 0.05$. Since by assumption a 0.95 prediction interval is to be computed, then $1 - \epsilon = .95$; or $\epsilon = 0.05$.

$t_{0.05} = 2.179$. This number is obtained from a table of values for the "t" distribution contained in G. W. Snedecor, Statistical Methods, op. cit., p. 65. The number 2.179 is found in the 0.05 column (corresponding to $\epsilon = 0.05$) on the 12 "degrees of freedom" row. In a regression analysis the term "degrees of freedom" means the sample size (N) minus the number of parameters in the regression equation. In our case N = 14 and there are 2 parameters (α and β) in the regression equation. Therefore degrees of freedom = 14-2 = 12.

$N = 14$. The sample size used as a basis for developing the estimating equation is 14.

$\hat{X} = 90$. By assumption.

$\overline{X}_2 = 37.9$. This is the value of the mean of the X_2's contained in Table 2.

$\Sigma(X_2 - \overline{X}_2)^2 = 14{,}673$. This quantity is best computed by using the following shortcut method:*

*The required numerical data are contained in Table 2.

$$\Sigma (X_2 - \bar{X}_2)^2 = \Sigma X_2^2 - (\Sigma X_2)^2/N$$

$$= 34,813 - 281,961/14$$

$$= 34,813 - 20,140$$

$$= 14,673.$$

Substituting the above data in equation (11), we have:

$$A = (25.01)(2.179)\sqrt{\frac{14}{14 - 2}\left[\frac{14 + 1}{14} + \frac{(90 - 38)^2}{14,673}\right]}$$

$$= (54.50)\sqrt{1.17\left[1.07 + \frac{2704}{14,673}\right]}$$

$$= (54.50)\sqrt{1.25 + (1.17)(0.184)}$$

$$= (54.50)\sqrt{1.465} = (54.50)(1.21)$$

$$= 65.95.$$

Therefore, for $\hat{X}_2 = 90$ the 95% prediction interval is:

$\hat{X}_1 \pm A$, or $172.5 \pm 66.0 = 106.5$ and 238.5.

This means that we have a subjective confidence of 95% that the interval 106.5 to 238.5 brackets the "true" or "population" value of X_1 corresponding to $\hat{X}_2 = 90$. It should be emphasized that a 95% prediction interval does <u>not</u> mean that the <u>probability</u> is 0.95 that the "true" value of X_1 lies within the interval. Rather it means that we are 95% "confident" (in a subjective sense) that this is the case. Statisticians call this <u>fiducial</u> probability, as opposed to a true

probability statement.*

Using the prediction interval procedure outlined above, we can compute 95% prediction intervals for $\hat{X}_1$ for numerous specified values of $\hat{X}_2$. The following are illustrative cases:

$\hat{X}_2$	$\hat{X}_1 \pm A$
10	$14.2 \pm 62.4 =$ -48.2 and 76.6
38	$69.6 \pm 61.0 =$ 8.6 and 130.6
60	$113.1 \pm 61.9 =$ 51.2 and 175.0
90	$172.5 \pm 66.0 =$ 106.5 and 238.5
120	$231.9 \pm 73.0 =$ 158.9 and 304.9
150	$291.2 \pm 81.8 =$ 209.4 and 373.0

Plotting these numbers on a scatter diagram and connecting the points, we obtain a 95% confidence band around the regression line. (See the heavy dashed lines in Fig. 5 on the next page.) In this case it is clear from the figure that the 95% confidence region is fairly wide, reflecting graphically a measure of the uncertainty associated with the estimating equation. This is rather typical of analyses based on small samples. The equation for the prediction interval is constructed so that the width of the interval is quite sensitive to variation in sample size when N is small. Sensitivity to small values of N is logical, since generalizations based on very small samples should be subject to greater uncertainty than those based on a larger data base.

*See Mood, op. cit., pp. 221-22.

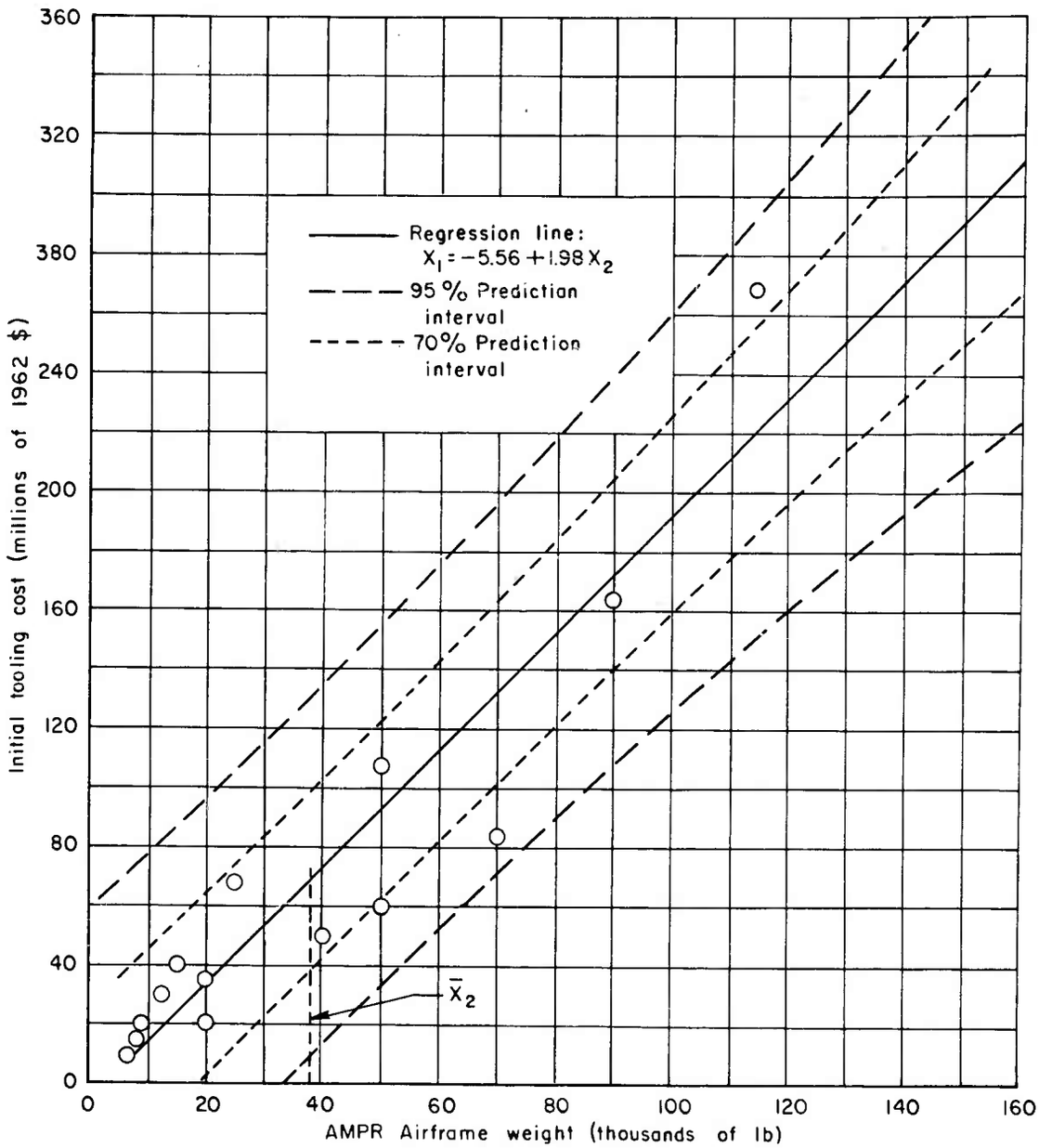

Fig.5— Initial tooling cost versus airframe weight

It should also be noted that because of the $(\hat{X}_2 - \bar{X}_2)^2$ term in equation (11) the prediction interval becomes wider as $\hat{X}_2$ is selected farther away from the mean of X_2 ($\bar{X}_2 = 38$) in the sample. Thus, for example, the prediction interval for $\hat{X}_2 = \bar{X}_2 = 38$ is:

$$69.6 \pm 61.0 = 8.6 \text{ and } 130.6;$$

while for $\hat{X}_2 = 200$ (which is considerably beyond the range of X_2 in the sample), the 95% prediction interval is:

$$390.2 \pm 99.7 = 290.5 \text{ and } 489.9.$$

The width of the interval in the latter case is over 1.6 times the width for $\hat{X}_2 = \bar{X}_2$:

$$\frac{(2)(99.7)}{(2)(61.0)} = \frac{199.4}{122.0} = 1.63.$$

This illustrates in a rough way how our "confidence" in the estimate decreases as we extrapolate beyond the range of the sample data -- something that we are almost always required to do in cost analysis of advanced weapon systems and forces.

The width of the prediction interval is also sensitive to the level of "confidence" specified. Up to now that level has been set at 95% (i.e., $\epsilon = 0.05$). Suppose that only a 70% level of confidence is desired ($\epsilon = 0.3$). The only thing that changes in the inputs used in the previous calculations is the value of t. Before, we used $t_{0.05} = 2.179$; now we use $t_{0.3} = 1.083$.* This, however, makes quite a difference

*Obtained from Table 3.8 in Snedecor, op. cit., p. 65; the 0.3 column and the d.f. = 12 row.

in the width of the prediction interval (now a 70% interval), as can be seen from the light dashed lines in Fig. 5 on page 23. Here, since our "confidence" is lower, the prediction interval can be narrower. For lower levels of confidence, the band would be even narrower. However, for a given level, the interval obtained by the prediction interval procedure will always be wider than an interval established on the basis of $\bar{S}$ alone.* This is apparent from Fig. 6 on the next page. The heavy dashed curves denote a 95% prediction region; the light dashed lines indicate a region established on the basis of $(\bar{S})(t_{0.05}) = (25.01)(2.179)$. Note that the two sets of boundaries are closest together where $\hat{X}_2 = \bar{X}_2 = 38$.

Up to this point, the discussion has been confined largely to statistical regression analyses -- developing an estimating (regression) equation and various measures of uncertainty pertaining to that equation. From an estimating point of view, this indeed is the most important part of the analysis.

There is, however, another form of statistical analysis called correlation analysis. Correlation analysis is concerned with developing an abstract measure of the degree of association between the dependent variable and the explanatory variable (or variables). In simple linear regression, the most commonly used measure of degree of

*But recall the point made previously: $\bar{S}$ can only be used to measure variations of X_1 in the sample -- not for describing the uncertainty of a predicted X_1.

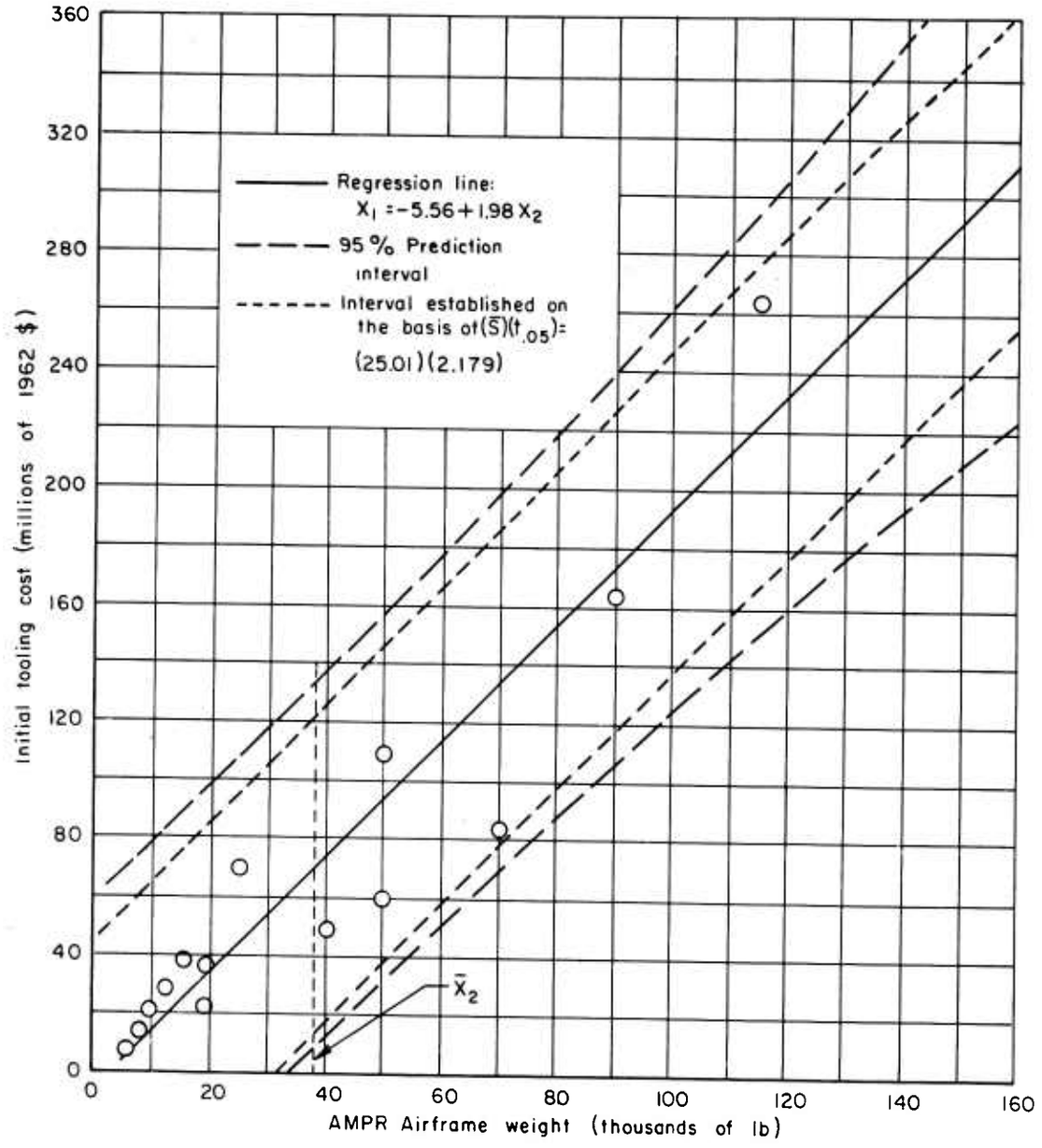

Fig. 6.— Initial tooling cost versus airframe weight

association is the correlation coefficient, denoted by r. The statistic r is constructed in such a way that it is bounded by the interval $-1 \leq r \leq +1$. The sign indicates whether the slope of the regression line is positive or negative -- i.e., whether the regression coefficient β is positive or negative. At the boundaries of the interval for r, we have the cases of perfect correlation: $r = +1$ (perfect positive correlation), $r = -1$ (perfect negative correlation). In these instances all of the sample points would lie exactly on the regression line. When there is no correlation between the variables whatsoever, $r = 0$.

While correlation is a somewhat different type of analysis from that discussed previously, it is nevertheless related in a definite way to regression analysis. In order to see this, let us return to the concepts of total variance, explained variance, and unexplained variance referred to earlier in the discussion of the standard error of estimate. Total variance pertains to the deviations of the sample X_1's from their mean, and is measured by:

$$\frac{\Sigma (X_1 - \bar{X}_1)^2}{N} \qquad (N = \text{sample size})$$

Explained variance refers to the deviations from $\bar{X}_1$ of the computed X_1 values (calculated from the regression equation) corresponding to the values of X_2 in the sample,* and is measured by:

*That is, for each value of X_2 in the sample, a corresponding value of the dependent variable (X_1) is computed from the regression equation $X_1 = \hat{\alpha} + \hat{\beta} X_2$.

$$\frac{\Sigma (X_{1_c} - \bar{X}_1)^2}{N}$$

Unexplained variance is derived from the deviations of the sample X_1 values from the computed values of X_1, and is measured by:

$$\frac{\Sigma (X_1 - X_{1_c})^2}{N}$$ *

Intuitively one would think that the standard error of estimate might somehow be derived from the unexplained variance. From our previous discussion, we recall that this is indeed the case. The standard error of estimate (unadjusted) is the square root of the unexplained variance.

Similarly, one would intuitively think that the correlation coefficient (r) might somehow be derived from the explained variance. The correlation coefficient is in fact related to the explained variance. It is defined as the square root of the proportion of total variance that is represented by the explained variance.** That is,

$$(12) \qquad r = \sqrt{\frac{\frac{\Sigma (X_{1_c} - \bar{X}_1)^2}{N}}{\frac{\Sigma (X_1 - \bar{X}_1)^2}{N}}} = \sqrt{\frac{\Sigma (X_{1_c} - \bar{X}_1)^2}{\Sigma (X_1 - \bar{X}_1)^2}}$$

We now see the interrelationship among r, S, and the regression

---

*A graphic portrayal of total, explained and unexplained variance is contained in Croxton and Cowden, op. cit., pp. 662-63.

**r^2 is sometimes referred to as the coefficient of determination.

equation. The regression equation is used to determine the computed X_1's, which are inputs to the calculation of both r and S. Also, since r^2 is defined as a proportion of total variance, r and S in a sense have an inverse relationship to one another.

Just as S had to be adjusted for sample size -- particularly so in the case of small samples -- r should also be corrected. (The formula for r -- equation (12) -- is for the unadjusted correlation coefficient.) In the case of simple linear correlation the value of r corrected for sample size is as follows:

$$(13) \qquad \bar{r} = \sqrt{\frac{r^2 (N - 1) - 1}{N - 2}} \; *$$

As is obvious from equation (13), the effect of the correction dampens out as N becomes large. For very small samples, as in our illustrative example, the correction should most certainly be made.

Returning to our illustrative example, we shall now compute the correlation coefficient. The formula for r as represented by equation (12) is rather cumbersome from a computational point of view. The following shortcut method is preferable:**

$$(14) \qquad r^2 = \frac{(\alpha \Sigma X_1 + \beta \Sigma X_1 X_2) - \bar{X}_1 \Sigma X_1}{\Sigma X_1^2 - \bar{X}_1 \Sigma X_1}$$

All of the data required for use in this equation have been computed

---

*See Croxton and Cowden, op. cit., p. 679.

**For discussion and derivation, see Croxton and Cowden, op. cit., p. 671, and Appendix B.

previously. We have calculated the regression coefficients to be: $\alpha = -5.5574$ and $\beta = 1.9789$. The mean of the X_1's and the summations are contained in Table 2 on page 10. Substituting these data in equation (14), we have:

$$r^2 = \frac{(-5.5574)(973) + (1.9789)(65,941) - (69.5)(973)}{132,589 - (69.5)(973)}$$

$$= \frac{-5,407.4 + 130,490.6 - 67,623.5}{132,589 - 67,623.5}$$

$$= \frac{57,459.7}{64,965.5} = 0.8845$$

$$r = \sqrt{0.8845} = 0.9405 \text{ (unadjusted)}$$

Using formula (13) to arrive at the correlation coefficient adjusted for sample size:

$$\bar{r}^2 = \frac{r^2 (N - 1) - 1}{N - 2}$$

$$= \frac{(0.8845)(14 - 1) - 1}{14 - 2}$$

$$= \frac{11.4985 - 1}{12} = \frac{10.4985}{12}$$

$$= 0.8749^*$$

$$\bar{r} = \sqrt{0.8749} = 0.9354$$

---

*Thus in this analysis, 87 per cent of the total variance in the sample X_1's is "explained" by the regression equation $X_1 = -5.56 + 1.98X_2$.

The fact that $\bar{r} = 0.94$, seems rather impressive. This represents a rather "high" correlation. But it is easy to be misled by high correlation coefficients. So much so, that in recent years there has been a trend away from the former emphasis on correlation analysis per se to regression analysis which stresses the derivation of structurally sound estimating relationships and of measures of the confidence that the user might have in the estimating equations.

In our illustrative example, it will be recalled, the measures of the unreliability of the estimating equation seemed rather high. The standard error of estimate in relation to the mean of the sample X_1's is high, and the confidence bands around the estimating equation would seem to be rather wide -- at least for certain applications. Yet the correlation coefficient turns out to be high, indicating a close "degree of association" between X_1 and X_2. This leads to an interesting question: How can the measure of correlation be so favorable and yet at the same time the measures of unreliability of the estimating equation tend to be unfavorable? Intuitively one can see why this might occur. The correlation coefficient is in a sense a measure of the average degree of association between the variables. However, it is conceivable that in an average sense the degree of association might be quite strong; but at the same time the dispersion or "spread" around the average may be fairly wide, thus leading to considerable uncertainty or unreliability of the estimating relationship.

This intuitive explanation appears plausible, but let us see if we can be more definitive. In order to do this, recall the previous

discussion of analysis of variance.* We have:

σ_u^2 = unexplained variance (S^2)

σ_e^2 = explained variance

σ_t^2 = total variance

Now r is not an absolute quantity, but rather it is based on a ratio of the explained to the total variance. To be precise:

$$r = \sqrt{\frac{\sigma_e^2}{\sigma_t^2}}$$

On the other hand, S is an absolute quantity: namely,

$$S = \sqrt{\sigma_u^2}$$

Here we have the key to the explanation that is being sought. Not infrequently the sample may be structured in such a way that the total variance (σ_t^2) will tend to be large.** Now even if the explained variance (σ_e^2) represents a high fraction of σ_t^2, and if σ_t^2 is large, there is still plenty of room for σ_u^2 (an absolute quantity) to be large; hence S can be large, especially in relation to the mean of the sample X_1's. In other words, the explanation hinges on the fact that

*For the sake of simplicity, in the discussion to follow we shall use the unadjusted S and r. This in no way affects the main line of argument.

**One circumstance that can lead to a large σ_t^2 is unequal distribution of observations within the sample. We shall illustrate this point later.

r is based on a ratio* while S is based on an absolute quantity; and that if total variance is large, S can still be large even though the quantity upon which it is based (S^2) represents a declining proportion of total variance as $r^2 = \sigma_e^2/\sigma_t^2$ increases.

The relationship between the standard error of estimate and the correlation coefficient may perhaps be seen more clearly from the following:

By definition, $\sigma_t^2 = \sigma_e^2 + \sigma_u^2$

Then, dividing through by σ_t^2, we have:

$$1 = \frac{\sigma_e^2}{\sigma_t^2} + \frac{\sigma_u^2}{\sigma_t^2},$$

or,

$$1 = r^2 + \frac{\sigma_u^2}{\sigma_t^2}$$

$$r^2 = 1 - \frac{\sigma_u^2}{\sigma_t^2}$$

$$r = \sqrt{1 - \frac{\sigma_u^2}{\sigma_t^2}} \qquad \left(\text{Recall } S = \sqrt{\sigma_u^2}\right)$$

To get S we may start with

$$r^2 = 1 - \frac{\sigma_u^2}{\sigma_t^2},$$

*Note that r is based on $r^2 = \sigma_e^2/\sigma_t^2$ and that if r^2 is a large fraction, r will be even larger since it is the square root of a number between zero and unity.

and derive

$$\sigma_t^2 r^2 = \sigma_t^2 - \sigma_u^2$$

$$\sigma_u^2 = \sigma_t^2 - \sigma_t^2 r^2$$

$$S = \sqrt{\sigma_t^2(1 - r^2)} = \sqrt{\sigma_u^2}$$

Returning to our illustrative example, the numerical values of the variances are:*

	Amount	Fraction of Total
Unexplained (σ_u^2)	536.1	0.12
Explained (σ_e^2)	4,104.3	0.88
Total (σ_t^2)	4,640.4	1.00

Let us examine these statistics. First, considering the total variance, it is not immediately obvious whether σ_t^2 is large or small. We can, however, readily determine that it is large. Taking the square root of $\sigma_t^2 = 4,640.4$, we obtain the standard deviation of the sample X_1's. This turns out to be $\sigma_t = 68.12$. The mean of the X_1's, it will be recalled, is $\bar{X}_1 = 69.5$. Therefore, in this case the standard deviation is essentially equal to the mean -- a situation that clearly indicates a wide dispersion of the variable under consideration.**

*Again, for simplicity we shall use unadjusted values of S and r.

**Recall that if a variable is normally distributed, an interval defined by the mean $\pm$ 1 standard deviation would include about 67 per cent of the cases.

What is the reason for the large standard deviation of the X_1's in our illustrative example? The answer is readily apparent from Fig. 7 on the next page. Here the sample observations on the scatter diagram are shown as deviations from the mean of the X_1's ($\bar{X}_1 = 69.5$). Note the uneven distribution of observations in the sample, with a few very large values on the high end. When the extremely large deviations of these few observations are squared -- as they must be in the calculation of the variance of the X_1's -- they have a magnified impact on the determination of σ_t^2. This is why it is desirable to have a more uniformly distributed sample; but often this is not possible, and we have to take what we can get. Incidentally, the uneven distribution of observations portrayed in Fig. 7, while having an "unfavorable" effect on σ_t^2, does not necessarily lower σ_e^2 and hence the correlation coefficient (r). Referring to the highest observation in Fig. 7, for example, the deviation of the observation from $\bar{X}_1$ (measured by the vertical line) contributes heavily to the magnitude of σ_t^2. However, the line AB (measuring the deviation of the <u>computed</u> X_1 from $\bar{X}_1$) enters into the calculation of <u>explained</u> variance (σ_e^2) and in effect "explains" most of the total variation. (Note again the "magnified" effects due to squaring these deviations.) In such cases, the correlation tends to become somewhat spurious, and because of this we should concentrate on regression analysis rather than correlation analysis.

Returning to the numerical values of the variances, it is clear that σ_e^2 is very large -- due in part to the "spuriousness" referred to earlier. In fact σ_e^2 "explains" 88 per cent of the total variance. Hence $r^2 = 0.88$, and $r = 0.94$. But as indicated previously, this is

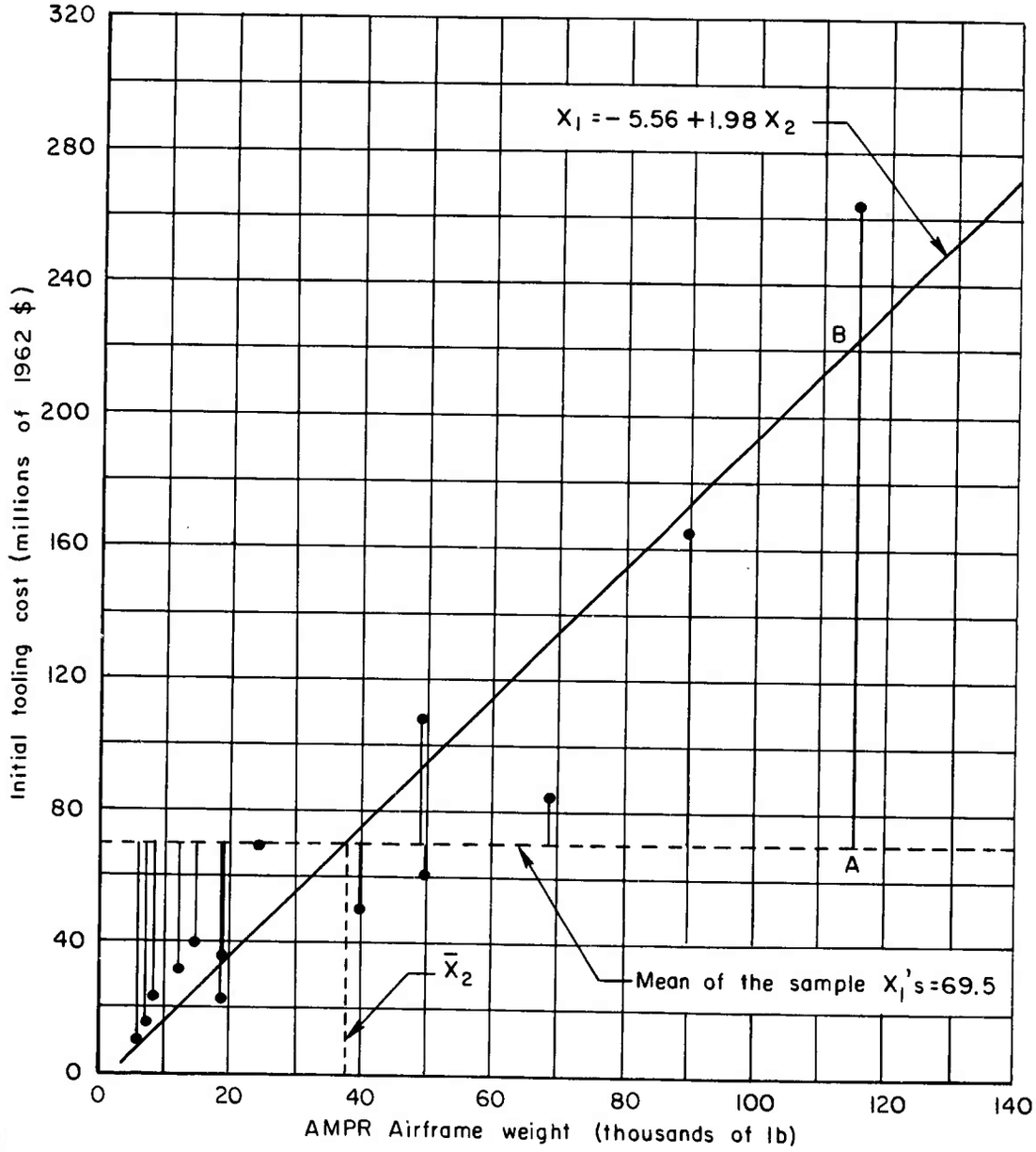

Fig. 7— Initial tooling cost versus airframe weight

misleading. For a more complete picture, we must turn to σ_u^2. Here the unexplained variance is only 12 per cent of total variance. However, in absolute terms it turns out to be fairly large: $\sigma_u^2 = 536$, which leads to $S = \sqrt{536} = 23$. The value of S per se does not mean very much. For perspective, we therefore relate it to the mean of the sample X_1's and obtain $S/\bar{X}_1 = 23/69.5 = 0.33$. A ratio this high does indeed indicate that S is quite large. Thus, in our numerical example we have a good illustration of the points made earlier in the general discussion: a large σ_t^2; σ_e^2 accounting for a large fraction of σ_t^2 (and hence a high correlation); and σ_u^2 accounting for a small fraction of σ_t^2, but at the same time being large in relation to the mean of the sample X_1's.

To illustrate these points still further, suppose that in our numerical example r^2 were even higher, say 0.95 (implying r = 0.97). The variance table would then be:

	Amount	Fraction of Total
Unexplained (σ_u^2)	232.0	0.05
Explained (σ_e^2)	4,408.4	0.95
Total (σ_t^2)	4,640.4	1.00

Here, $S = \sqrt{232} = 15.2$; and $S/\bar{X}_1 = 15.2/69.5 = 0.22$. Thus even if σ_u^2 accounted for only 5 per cent of σ_t^2, the ratio of $S/\bar{X}_1$ would be 0.22 -- still fairly high.*

*At this point the students will be required to do a simple linear regression analysis. (See Appendix B.).

IV. A CURVILINEAR ANALYSIS: LOGARITHMIC REGRESSION

Up to this point the analysis has been confined to simple linear regression. While a first examination of the scatter diagram of X_1 vs. X_2 indicates that a linear relationship might be as good as anything else, it still cannot be concluded definitely that some type of non-linear relationship might not be preferable. We shall now explore this possibility.

One type of non-linear relationship that is very frequently used is of the form:

$$X_1 = \alpha X_2^{\beta} \tag{15}$$

Equation (15), however, is difficult to deal with statistically; so usually we make a logarithmic transformation of the variables, obtaining an equation which is linear in the logarithms of the variables:

$$\log X_1 = \log \alpha + \beta \log X_2 \tag{16}$$

The procedure here is to conduct the statistical analysis in terms of the logarithms of the variables -- obtaining estimates of $\log \alpha$ and β from a least squares fit of equation (16) and then transforming back to the original data and to equation (15). This approach has several advantages, the most important ones being:

(1) We can proceed almost identically to the simple linear regression case.

(2) No additional degrees of freedom are lost* -- an important consideration when the sample size is small.

The first step is to take the original data for X_1 and X_2 contained in Table 2 and convert these data to logarithms. The cross product and the squares are then computed, and the summations are derived. The results of these calculations are presented in Table 3. We can now proceed to a simple linear regression analysis of the data in logarithmic form.** This means that a linear regression equation is derived, such that the sum of squares of the logarithms of the variables around the regression line is at a minimum.

The "normal equations" for estimating the regression coefficients are the same form as before:

$$(17) \qquad \Sigma \log X_1 = N \log\alpha + \beta \Sigma \log X_2$$

$$(18) \qquad \Sigma\left[(\log X_1)(\log X_2)\right] = \log\alpha \Sigma \log X_2 + \beta \Sigma (\log X_2)^2$$

Substituting the required summations contained in Table 3 into equations (17) and (18), we obtain:

$$(19) \qquad 23.2383 = 14 \log\alpha + 19.8177\beta$$

$$(20) \qquad 34.9241 = 19.8177 \log\alpha + 30.1372\beta$$

Solving equations (19) and (20) simultaneously (using the same procedure

---

*Recall that in a regression analysis, degrees of freedom means the sample size minus the number of parameters in the estimating equation. Since in logarithmic regression there are only two parameters in the estimating equation, the number of degrees of freedom is the same as for simple linear regression: N - 2.

**See Croxton and Cowden, op. cit., pp. 694-99.

Table 3

DATA FOR LOG-LINEAR REGRESSION ANALYSIS OF INITIAL TOOLING COST AND AIRFRAME WEIGHT

Aircraft Type	Log X_1	Log X_2	(Log X_1)(Log X_2)	(Log X_1)^2	(Log X_2)^2
F-1	0.9031	0.8451	0.7632	0.8156	0.7142
F-2	1.1761	0.9031	1.0621	1.3832	0.8156
F-3	1.3010	0.9542	1.2414	1.6926	0.9105
F-4	1.6021	1.1761	1.8842	2.5667	1.3832
F-5	1.4771	1.0792	1.5941	2.1818	1.1647
F-6	1.5441	1.3010	2.0089	2.3842	1.6926
F-7	1.8451	1.3979	2.5793	3.4044	1.9541
B-1	1.6990	1.6021	2.7220	2.8866	2.5667
B-2	2.4232	2.0607	4.9935	5.8719	4.2465
B-3	2.0414	1.6990	3.4683	4.1673	2.8866
B-4	1.9294	1.8451	3.5599	3.7226	3.4044
B-5	1.7782	1.6990	3.0212	3.1620	2.8866
B-6	1.3010	1.3010	1.6926	1.6926	1.6926
B-7	2.2175	1.9542	4.3334	4.9173	3.8189
Totals	23.2383	19.8177	34.9241	40.8488	30.1372

SOURCE: X_1 and X_2 data contained in Table 2, converted to logarithms.

$$\frac{\Sigma \log X_1}{N} = \frac{23.2383}{14} = 1.6599$$

$$\frac{\Sigma \log X_2}{N} = \frac{19.8177}{14} = 1.4156$$

as before), the estimates of $\log \alpha$ and β are found to be:

$$\log \alpha = 0.281824$$

$$\beta = 0.973516^{*}$$

The regression equation for the logarithms of the variables is, therefore:

$$(21) \qquad \log X_1 = 0.2818 + 0.9735 \log X_2$$

Equation (21) is plotted on the scatter diagram contained in Fig. 8-A (the solid line). Note that here the original values (arithmetic form) of X_1 and X_2 are plotted on a chart having logarithmic scales on both axes (a "log-log" chart). This is exactly equivalent to plotting the logarithms of the variables on an arithmetic chart. (See Fig. 8-B.)

The standard error of estimate is computed as before:

$$\bar{s}^2_{\log} = \frac{\Sigma(\log X_1)^2 - \log\alpha\Sigma\log X_1 - \beta\Sigma[(\log X_1)(\log X_2)]}{N - 2}$$

$$= \frac{40.8488 - (0.281824)(23.2383) - (0.973516)(34.9241)}{14 - 2}$$

$$= \frac{40.8488 - 6.5491 - 33.9992}{12}$$

$$= \frac{0.3005}{12} = 0.02504$$

$$\bar{s}_{\log} = \sqrt{0.02504} = 0.1582.$$

---

*Substituting these values for $\log \alpha$ and β in equation (19), we obtain a check as follows:

$$23.2383 = (14)(0.281824) + (19.8177)(0.973516)$$
$$23.2383 = 3.9455 + 19.2928$$
$$23.2383 = 23.2383.$$

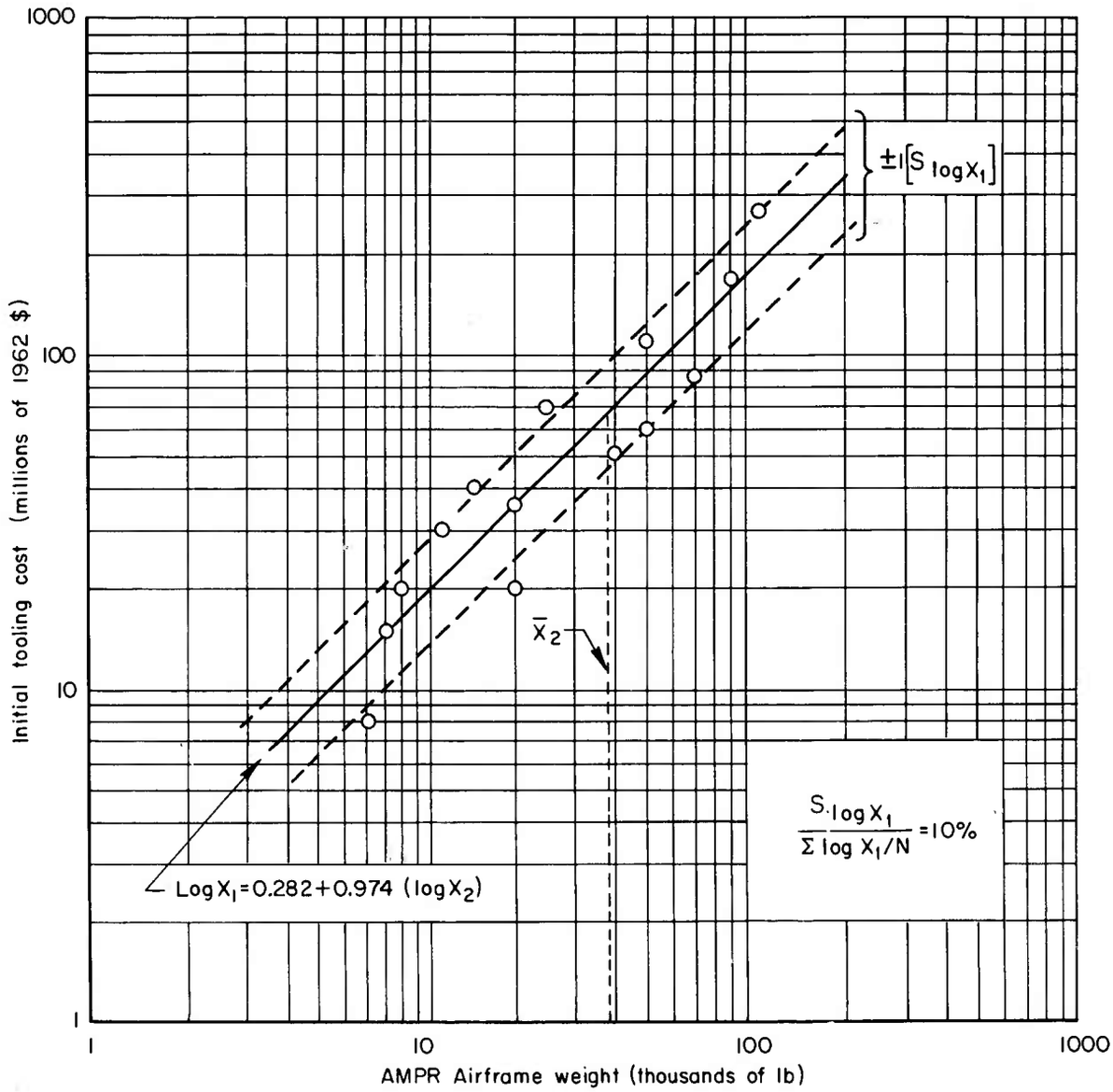

Fig. 8a—Initial tooling cost versus airframe weight

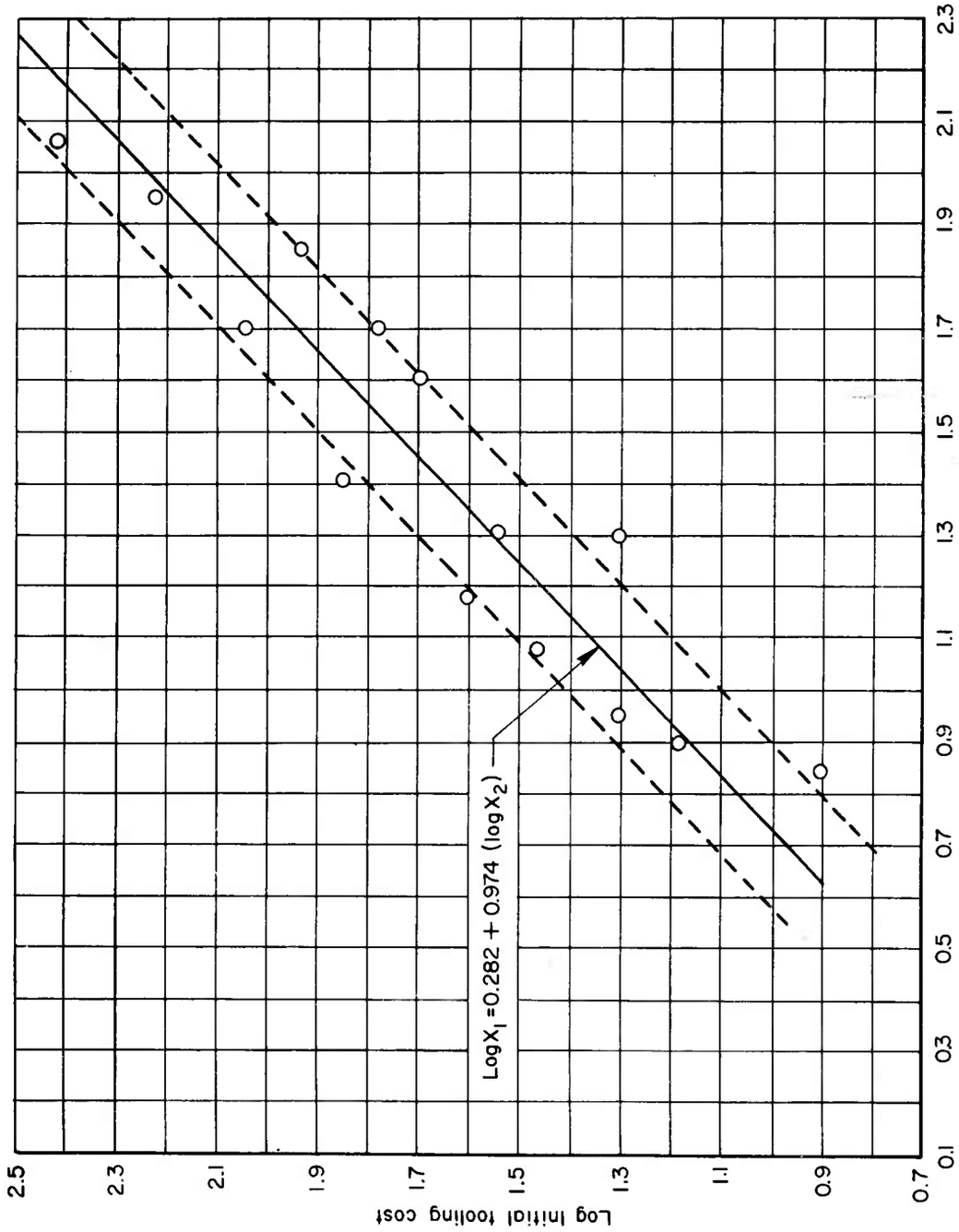

Fig.8b— Initial tooling cost versus airframe weight

In Figs. 8-A and 8-B, the dashed lines indicate a band representing $\pm 1\ \bar{S}_{\log}$ around the regression line $\log X_1 = 0.282 + 0.974 \log X_2$.

For perspective, the value of $\bar{S}_{\log}$ may be related to the mean of the $\log X_1$'s in the sample:

$$\frac{\bar{S}_{\log}}{\Sigma \log X_1 / N} = \frac{0.1582}{1.6599} = 0.095.$$

At this point it would appear that things have improved markedly over the simple linear regression case. The picture portrayed in Figs. 8-A and 8-B suggests an excellent "fit" to the data. Also, the standard error of estimate in relation to the mean of the $\log X_1$'s is substantially lower than in the simple linear regression example: 10 per cent as compared with 36 per cent.

But this is not the whole story. Up to now the analysis has dealt with the <u>logarithms</u> of the data. The analyst, however, is not interested in estimating $\log X_1$ for a given value of $\log X_2$; rather he is interested in making estimates in terms of the <u>original</u> data. We therefore have to transform the logarithmic analysis back to an arithmetic form. When this transformation is made, the log-linear estimating equation

$$\log X_1 = 0.2818 + 0.9735 \log X_2$$

becomes

$$X_1 = 1.9135\ X_2^{0.9735}, \tag{22}$$

where 1.9135 is the anti-log of $\log \alpha = 0.2818$. Equation (22) is

plotted on the scatter diagram contained in Fig. 9 (the solid line). It should be noted that in this case equation (22) plots as a straight line over the range of X_2 shown in Fig. 9. Since the exponent of X_2 is so close to unity, the curvilinearity implied by the form of (22) does not show up. For all practical purposes, equation (22) plots X_1 as a linear homogeneous function of X_2. Note also that the regression line does not appear to be a particularly good "fit" to the original data -- certainly no better than the simple linear estimating equation obtained previously.

To gain further insight, let us turn to the standard error of estimate and compute a $\pm 1\ \bar{S}$ band about the regression line. Again, we must transform the logarithmic analysis into an arithmetic one. This may be done in two ways. One way is to carry through the computation in terms of logarithms and convert to the original data at the very end. As an illustration, assume that the analyst wants to compute an estimate of X_1 for $X_2 = 100$. The logarithm of 100 is 2. We have, then:

$$\begin{aligned} \log X_1 &= 0.2818 + 0.9735(2) \\ &= 0.2818 + 1.9470 \\ &= 2.2288 \end{aligned}$$

$$\log X_1 \pm \bar{S}_{\log} = 2.2288 \pm 0.1582 = 2.0706 \text{ and } 2.3870.$$

These latter two numbers are converted into arithmetic terms by taking the anti-logarithms:

$$\text{anti-log } 2.3870 = 243.8$$

$$\text{anti-log } 2.0706 = 117.7$$

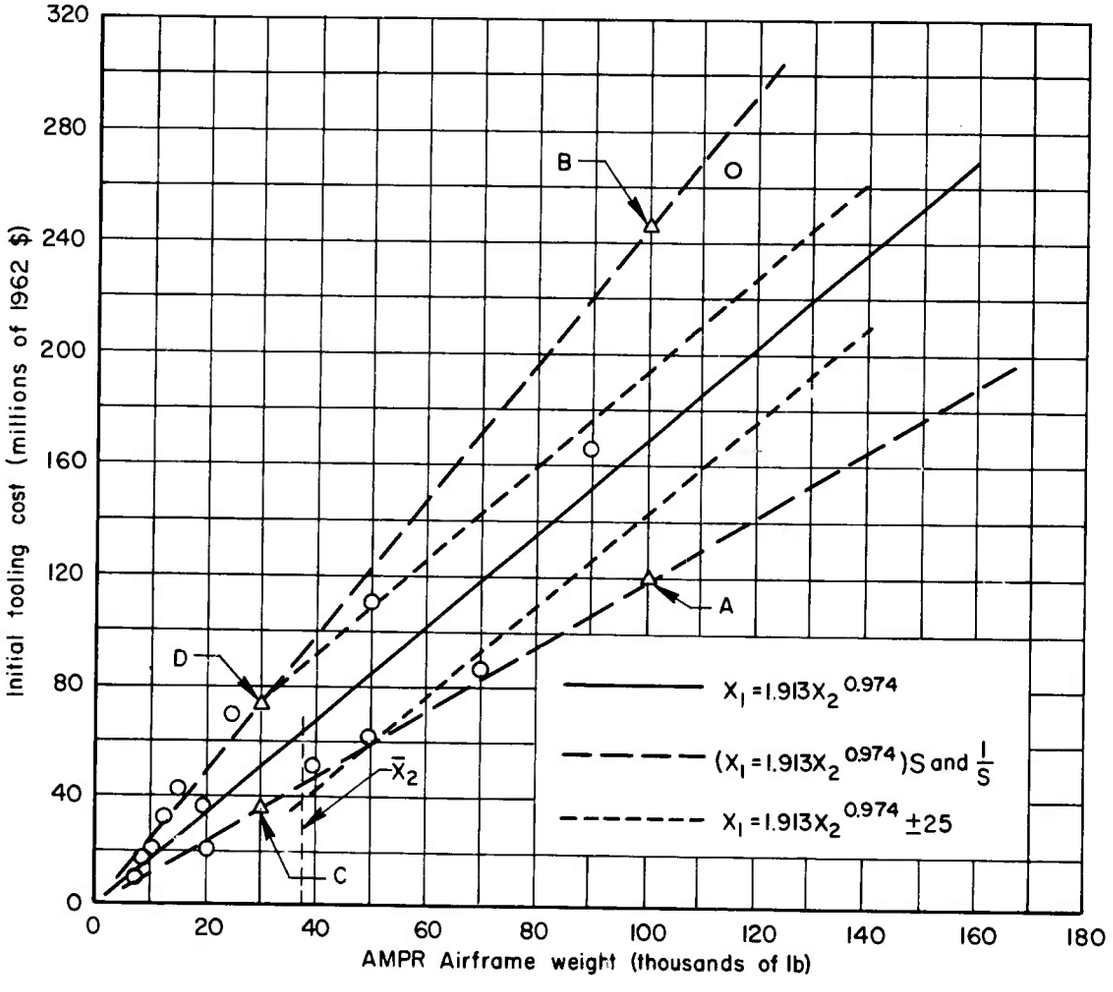

Fig. 9 — Initial tooling cost versus airframe weight

The $\pm \bar{S}$ interval for $X_2 = 100$ is, therefore, 118 to 244. (See points A and B in Fig. 9.)

Another approach is to convert from a logarithmic to an arithmetic approach immediately. In the previous method we had

$$\log X_1 \pm \bar{S}_{\log};$$

but in terms of the original data, this transforms into (anti-log X_1) (anti-log $\bar{S}_{\log}$) for the "plus" case, and (anti-log X_1) $\div$ (anti-log $\bar{S}_{\log}$) for the "minus" case.* First we need the anti-log of $\bar{S}_{\log}$, which is 1.4397. In the previous example we found that for $X_2 = 100$, $\log X_1 = 2.2288$. The anti-log of 2.2288 gives us $X_1 = 169.34$. The $\pm \bar{S}$ interval is, therefore:

$$(169.34)(1.4397) = 243.8$$

$$(169.34)(1/1.4397) = 117.6,$$

which is the same as that obtained by the first method.

As another example, let us obtain the $\pm \bar{S}$ interval for $X_2 = 30$. From the regression line in Fig. 9, we read off $X_1 \cong 51$ when $X_2 = 30$. The interval is:

$$(51)(1.4397) = 73$$

$$(51)(1/1.4397) = 35,$$

which when plotted gives points C and D in Fig. 9. Connecting points A and C for the lower bound and points B and D for the upper bound, we obtain the $\pm \bar{S}$ interval around the regression equation $X_1 = 1.913X_2^{0.974}$.

---

*Recall that addition of logarithms is equivalent to multiplication in arithmetic terms, and subtraction of logarithms is equivalent to division in arithmetic terms.

We now have a much different picture than that indicated in Figs. 8-A and 8-B for the logarithmic analysis. In Fig. 9 the $\pm \bar{S}$ interval is an ever-widening one defined in terms of linear homogeneous functions of X_2, with slope 1.44 for the upper bound and slope $1/1.44 = 0.69$ for the lower bound. (See the heavy dashed lines in Fig. 9.)

Recall that in our simple linear regression analysis in Section III, $\bar{S} = 25$. If we lay off ± 25 around the regression line $X_1 = 1.913X_2^{0.974}$, the results are the light dashed lines in Fig. 9. Here it is interesting to note that at approximately the mean of the sample X_1's ($\bar{X}_1 = 38$) the two $\pm \bar{S}$ intervals are the same width. For ranges of $X_2 < 38$, the ± 25 interval is the larger; and for $X_2 > 38$, the ± 25 interval is the smaller of the two. Thus, while for the very low range of values for X_2 we might prefer using $X_1 = 1.913X_2^{0.974}$ as an estimating equation, we would not prefer it over the simple linear regression equation for the majority of the range of X_2 in the sample. We conclude, therefore, that in general $X_1 = 1.913X_2^{0.974}$ offers no improvement over $X_1 = -5.56 + 1.98\ X_2$.

The logarithmic example contained in this section illustrates a point that is often forgotten. A logarithmic transformation of the variables has a tendency to compress and shape the original data in such a way that a statistical fit to the logarithms "looks good." However, as pointed out previously, we are not interested in estimating the logarithms. Very often when the logarithmic analysis is transformed back into terms of the original data, the results do not appear so impressive -- as was the case in our example. In sum, logarithmic transformations can be tricky and misleading. We must be cautious when using them.

V. A CURVILINEAR ANALYSIS: SECOND DEGREE EQUATION

We have just seen that for our illustrative example a logarithmic regression does not seem to offer any improvement over the simple linear regression case. In this section another type of curvilinear regression analysis will be attempted. Here, a second degree equation will be used.

The estimating equation is of the form:

$$(23) \qquad X_1 = \alpha + \beta_1 X_2 + \beta_2 X_2^2$$

In this case three parameters must be estimated: α, β_1, and β_2.* Instead of two "normal equations" we now must have three. They are:**

$$(24) \qquad \Sigma X_1 = \alpha N + \beta_1 \Sigma X_2 + \beta_2 \Sigma X_2^2$$

$$(25) \qquad \Sigma X_1 X_2 = \alpha \Sigma X_2 + \beta_1 \Sigma X_2^2 + \beta_2 \Sigma X_2^3$$

$$(26) \qquad \Sigma X_1 X_2^2 = \alpha \Sigma X_2^2 + \beta_1 \Sigma X_2^3 + \beta_2 \Sigma X_2^4$$

Most of the summation data required for these equations are contained in Table 2: namely,

$$\Sigma X_1 = 973$$

$$\Sigma X_2 = 531$$

---

*Notice that by adding the variable X_2^2, an additional degree of freedom is lost. In the simple linear regression case, degrees of freedom were N - 2 = 14 - 2 = 12. Now we have N - 3 = 14 - 3 = 11 degrees of freedom.

**See Croxton and Cowden, op. cit., p. 706.

$$\Sigma X_1 X_2 = 65,941$$

$$\Sigma X_2^2 = 34,813$$

$$\Sigma X_1^2 = 132,589$$

However, the following supplementary data are needed: $\Sigma X_1 X_2^2$, ΣX_2^3, and ΣX_2^4. The data are calculated and presented in Table 4.

Substituting these summations into equations (24), (25), and (26):

(27) $$973 = \alpha 14 + \beta_1 531 + \beta_2 34,813$$

(28) $$65,941 = \alpha 531 + \beta_1 34,813 + \beta_2 2,945,187$$

(29) $$5,844,667 = \alpha 34,813 + \beta_1 2,945,187 + \beta_2 280,375,669$$

These equations may be solved simultaneously by repeated use of the same technique that was used in Section III. Here, we shall take (27) and (28) together and eliminate α; do the same thing for (28) and (29); take the resulting two equations in β_1 and β_2 and eliminate β_1; solve the result for β_2; and then substitute back in previous equations to get α and β_1.

Following this procedure, the calculations are as follows:

Ratio of the coefficients of α in equation (27) and (28) = $\frac{531}{14} = 37.928571$

Multiplying equation (27) by 37.928571 and subtracting the resulting equation from (28):

$$65,941 = \alpha 531 + \beta_1 34,813 + \beta_2 2,945,187$$

Table 4

SUPPLEMENTARY DATA NEEDED FOR SECOND DEGREE REGRESSION ANALYSIS

Observation	$X_1X_2^2$	X_2^3	X_2^4
F-1	392	343	2,401
F-2	960	512	4,096
F-3	1,620	729	6,561
F-4	9,000	3,375	50,625
F-5	4,320	1,728	20,736
F-6	14,000	8,000	160,000
F-7	43,750	15,625	390,625
B-1	80,000	64,000	2,560,000
B-2	3,504,625	1,520,875	174,900,625
B-3	275,000	125,000	6,250,000
B-4	416,500	343,000	24,010,000
B-5	150,000	125,000	6,250,000
B-6	8,000	8,000	160,000
B-7	1,336,500	729,000	65,610,000
Total	5,844,667	2,945,187	280,375,669

$$36{,}905 = \alpha 531 + \beta_1 20{,}140 + \beta_2 1{,}320{,}407$$

(30) $$29{,}036 = \beta_1 14{,}673 + \beta_2 1{,}624{,}780$$

Ratio of the coefficients of α in equations (28) and (29) = $\frac{34{,}813}{531} = 65.561205$.

Multiplying equation (28) by 65.561205 and subtracting the result from equation (29):

$$5{,}844{,}667 = \alpha 34{,}813 + \beta_1 2{,}945{,}187 + \beta_2 280{,}375{,}669$$

$$4{,}323{,}171 = \alpha 34{,}813 + \beta_1 2{,}282{,}382 + \beta_2 193{,}090{,}009$$

(31) $$1{,}521{,}496 = \beta_1 662{,}805 + \beta_2 87{,}285{,}660$$

Now taking equations (30) and (31), we eliminate β_1 by multiplying equation (30) by 662,805/14,673 = 45.171744 and subtracting the result from equation (31):

$$1{,}521{,}496 = \beta_1 662{,}805 + \beta_2 87{,}285{,}660$$

$$1{,}311{,}607 = \beta_1 662{,}805 + \beta_2 73{,}394{,}146$$

$$209{,}889 = \beta_2 13{,}891{,}514$$

$$\beta_2 = \frac{209{,}889}{13{,}891{,}514} = \underline{0.015109}$$

Substituting $\beta_2 = 0.015109$ in equation (30) and solving for β_1:

$$29{,}036 = \beta_1 14{,}673 + (1{,}624{,}780)(.015109)$$

$$29,036 = \beta_1 14,673 + 24,549$$

$$\beta_1 14,673 = 4,487$$

$$\beta_1 = \underline{\underline{0.305800}}$$

Substituting $\beta_1 = 0.015109$ and $\beta_2 = 0.305800$ in equation (27) and solving for α:

$$973 = \alpha 14 + (.305800)(531) + (0.015109)(34,813)$$

$$973 = \alpha 14 + 162 + 526$$

$$\alpha 14 = 285$$

$$\alpha = \underline{\underline{20.357143}}$$

Checking the computations by substituting the derived values of α, β_1, and β_2 into equation (28):

$$65,941 = (20.357143)(531) + (0.305800)(34,813) + (0.015109)(2,945,187)$$

$$65,941 = 10,810 + 10,646 + 44,499$$

$$65,941 \cong 65,955$$

Taking the estimates of α, β_1 and β_2 derived above, the estimating equation becomes:

(32) $$X_1 = 20.36 + 0.3058X_2 + 0.0151X_2^2$$

This equation is plotted on the scatter diagram contained in Fig. 10 on the next page (the solid line).

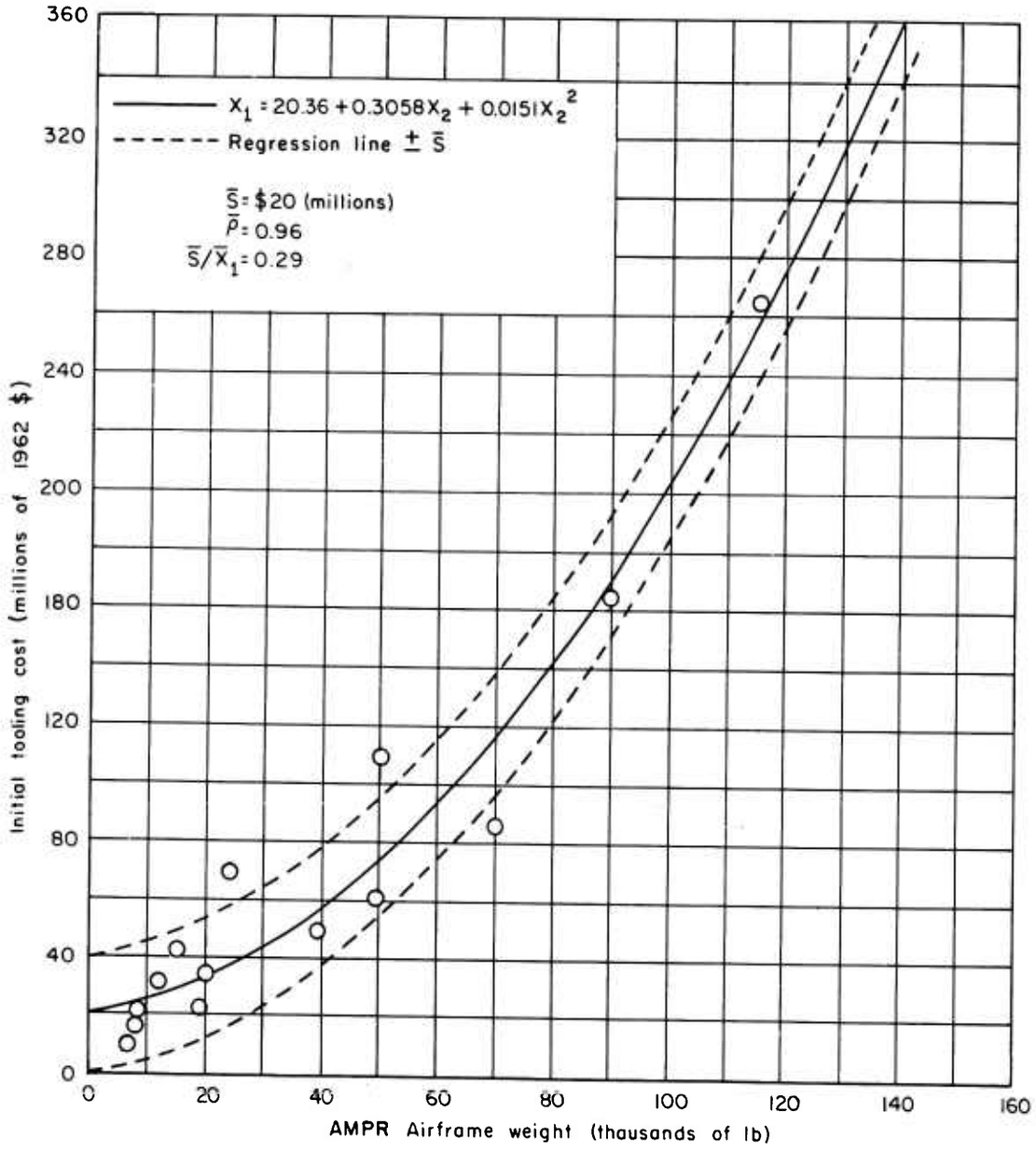

Fig. 10—Initial tooling cost versus airframe weight
(2nd degree case)

The standard error of estimate is calculated as before, except here we have to add a term for β_2 and take into account the loss of the additional degree of freedom. The formula is:

$$(33) \qquad \bar{S}^2 = \frac{\Sigma X_1^2 - (\alpha \Sigma X_1 + \beta_1 \Sigma X_1 X_2 + \beta_2 \Sigma X_1 X_2^2)}{N - 3}$$

$$= \frac{132,589 - (20.3571)(973) - (0.3058)(65,941) - (0.0151)(5,844,667)}{14 - 3}$$

$$= \frac{132,589 - 19,807 - 20,165 - 88,254}{11}$$

$$= \frac{4,363}{11} = 396.64$$

$$\bar{S} = \sqrt{396.64} = \$19.92 \text{ (Millions)}$$

Relating $\bar{S} = 19.92$ to the mean of the sample X_1's:

$$\frac{\bar{S}}{\bar{X}_1} = \frac{19.92}{69.5} = 0.287$$

An area bounded by $\pm 1\bar{S}$ around the regression line is presented in Fig. 10 (the dashed lines).

As in the simple linear regression case, a prediction interval may be calculated for a value of X_1 obtained from the estimating equation for specified values of X_2 and X_2^2. For a second degree regression, however, the calculation is somewhat more complicated. Since the computational procedure required here is the same as that for multiple regression analysis, we shall defer the subject of prediction intervals until the following section (Section VI) on multiple regression analysis.

We now turn to calculation of the measures of correlation. In curvilinear analysis, the coefficient of curvilinear correlation is usually referred to as the index of correlation and is denoted by the

symbol ρ. ρ^2 is called the index of determination. The formula for ρ^2 is:

$$(34) \qquad \rho^2 = \frac{\Sigma X_{1_c}^2 - \bar{X}_1 \Sigma X_1}{\Sigma X_1^2 - \bar{X}_1 \Sigma X_1},$$

where

$$(35) \qquad \Sigma X_{1_c}^2 = \alpha \Sigma X_1 + \beta_1 \Sigma X_1 X_2 + \beta_2 \Sigma X_1 X_2^2.$$

Equation (34) gives the unadjusted ρ^2. Particularly in the case of small samples, ρ^2 should be adjusted for degrees of freedom. The following formula may be used for this purpose:

$$(36) \qquad \bar{\rho}^2 = \frac{\rho^2(N - 1) - (m - 1)}{N - m},$$

where m is the number of coefficients in the regression equation. (m = 3 in the case of second degree regression.)

Substituting the required data in equations (34) and (35), we have:

$$\Sigma X_{1_c}^2 = (20.3571)(973) + (0.3058)(65,941) + (0.0151)(5,844,667)$$

$$= 19,807 + 20,165 + 88,254$$

$$= 128,226.$$

$$\rho^2 = \frac{128,226 - (69.50)(973)}{132,589 - (69.50)(973)}$$

$$= \frac{128,226 - 67,624}{132,589 - 67,624}$$

$$= \frac{60,602}{64,965} = 0.9328$$

$$\bar{\rho}^2 = \frac{(0.9328)(14 - 1) - (3 - 1)}{14 - 3}$$

$$= \frac{12.1264 - 2}{11}$$

$$= \frac{10.1264}{11} = 0.9206$$

$$\bar{\rho} = \sqrt{0.9206} = 0.9595.$$

We may now compare the results of the statistical analysis for the second degree regression case with those obtained for the simple linear regression example:*

	Simple Linear Regression	Second Degree Regression
Standard error of estimate	$25 (million)	$20 (million)
Coefficient of Variation ($\bar{S}/\bar{X}_1$)	0.36	0.29
Coefficient (index) of determination	0.87	0.92
Coefficient (index) of correlation	0.94	0.96

From these data it would appear that the second degree regression offers a considerable improvement over the simple linear case. The standard error of estimate is reduced by $5 million, the coefficient of variation is lower by 7 percentage points, and the percentage of "explained" variation is higher by 5 percentage points. Also, the regression curve in Fig. 10 would appear to be a very good "fit" to the sample data.

*All measures included here are adjusted for degrees of freedom.

The real question, however, is whether the improvement is significant in a statistical sense. It must not be forgotten that in our illustrative examples we are dealing with a sample of data, and a very small sample indeed. It is conceivable, therefore, that the differences in the statistical measures presented above could be attributable purely to sampling error, and that in the "universe" or "population" we were attempting to describe, there is in reality no improvement in going from a linear to a second degree estimating equation. If this were to be so, then the "improvement" we have observed would not be regarded as "significant" in a statistical sense.

In order to resolve a question like this, the analyst must resort to a statistical testing procedure -- a rather complex subject, the details of which are beyond the scope of the present discussion. Basically what is involved in a statistical test is to set up the hypothesis that the observed differences are in effect non-existent, and then let the testing procedure indicate whether the hypothesis is accepted or rejected at some pre-specified level of probability. In a specific case, for example, we might have two statistical measures x_1 and x_2, with a difference of Δx. The statistical test might indicate that the chances are very small that two samples drawn from the assumed population would have statistical measures leading to a difference as large as Δx. In other words it would seem highly unlikely that the observed difference could be attributable to sampling variation. If this were the case, we would conclude that the difference between x_1 and x_2 is significant, and that therefore the hypothesis that $x_1 = x_2$ is rejected.

In our present example, the author did conduct such a test.

Specifically, a statistical test was made to determine whether the difference in explained variation (0.87 for simple linear regression vs. 0.92 for second degree regression) is significant. Or, stating the problem another way, a test was made to determine whether the incremental increase in explained variance associated with the addition of the variable X_2^2 is significant. The results of the test indicate that the chances are extremely small (less than 1 in 20 in this case) that the observed difference could be attributable to sampling error alone.* We conclude, therefore, that the statistical results of the curvilinear regression are significantly better than those for the simple linear case.**

---

*For a discussion of the testing procedure used, see Croxton and Cowden, op. cit., pp. 710-12.

**At this point the students will be required to do a curvilinear regression analysis. (See Appendix C.)

VI. A MULTIPLE REGRESSION ANALYSIS

In Section V the simple linear regression example was extended by introducing the variable X_2^2 into the estimating equation, resulting in a curvilinear regression analysis. We shall now go back to the simple linear case and consider adding a new variable to the regression equation. This takes us into the realm of multiple regression analysis. Since here we shall introduce the new variable in a linear fashion, the analysis will represent one class of multivariate analysis: multiple linear regression.

The first question concerns which new variable to introduce into the regression equation. As indicated in Section II, in addition to X_2 (airframe weight), we have data on two other variables: X_3 (maximum speed) and X_4 (combat radius). At this point a technical consideration must be raised. The multiple regression model that is used in the analysis to follow, postulates that the explanatory variables be non-correlated. We must therefore examine the relationship between X_2 and X_3, and X_2 and X_4. While there are statistical techniques for testing whether or not a significant correlation exists between two explanatory variables, a more simplified procedure will be used here. We shall merely examine the scatter diagrams for X_2 vs. X_3 and X_3 vs. X_4. These are presented in Figs. 11 and 12. From Fig. 11 it is clear that a considerable amount of correlation exists between X_2 and X_4. From Fig. 12 it would seem that while some degree of association may exist between X_2 and X_3, the correlation is certainly not very great. Therefore, X_3 will be chosen as the additional variable to be introduced into the estimating equation which will be of the form:

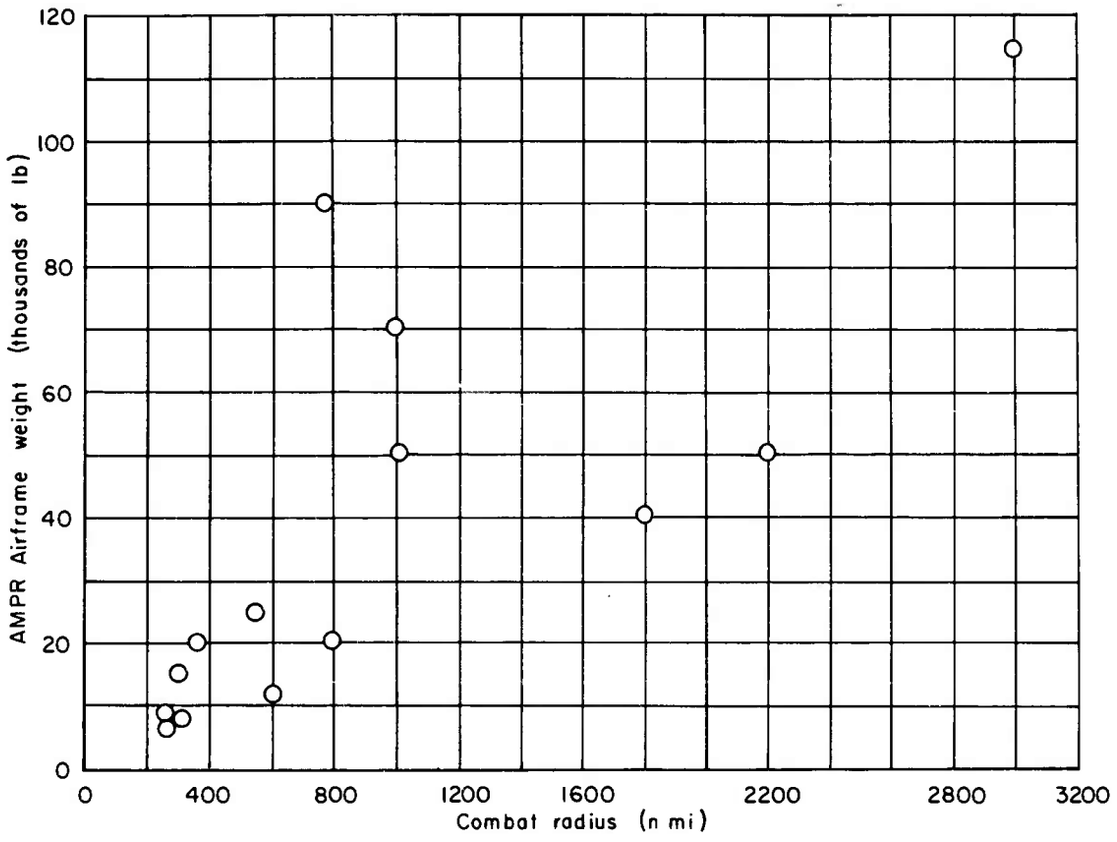

Fig. 11— Airframe weight versus combat radius

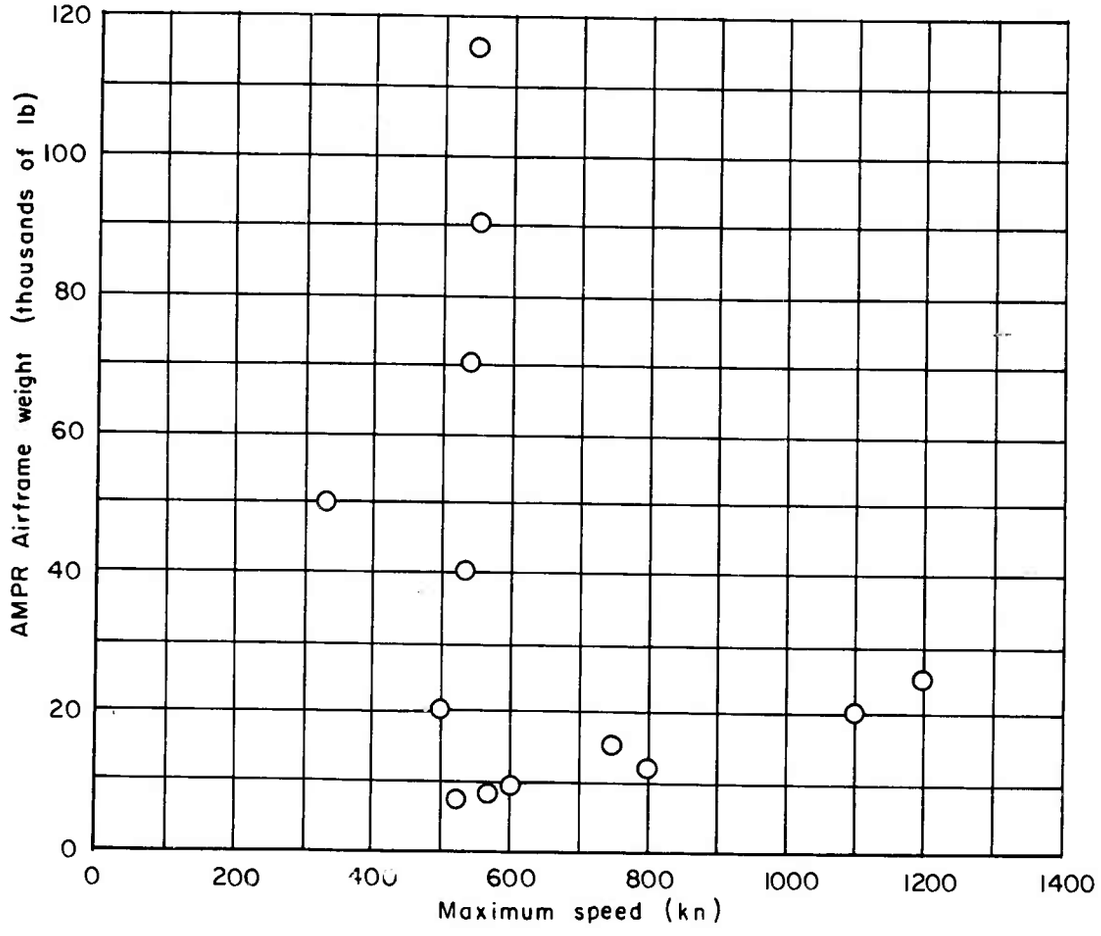

Fig.12 — Airframe weight versus maximum speed

$$(37) \qquad X_1 = \alpha + \beta_{12.3}X_2 + \beta_{13.2}X_3$$

The normal equations required to obtain estimates of the parameters α, $\beta_{12.3}$, and $\beta_{13.2}$ are:*

$$(38) \qquad \Sigma X_1 = N\alpha + \beta_{12.3}\Sigma X_2 + \beta_{13.2}\Sigma X_3$$

$$(39) \qquad \Sigma X_1X_2 = \alpha\Sigma X_2 + \beta_{12.3}X_2^2 + \beta_{13.2}\Sigma X_2X_3$$

$$(40) \qquad \Sigma X_1X_3 = \alpha\Sigma X_3 + \beta_{12.3}\Sigma X_2X_3 + \beta_{13.2}\Sigma X_3^2$$

These equations may be solved simultaneously, using the same procedure outlined in Section V for the second degree regression case. However, the three normal equations may be reduced to two if the analysis is conducted in terms of deviations from the means of the variables. Denoting deviations by small x,

$$x_i = X_i - \bar{X}_i,$$

the normal equations become:

$$(41) \qquad \Sigma x_1x_2 = \beta_{12.3}\Sigma x_2^2 + \beta_{13.2}\Sigma x_2x_3$$

$$(42) \qquad \Sigma x_1x_3 = \beta_{12.3}\Sigma x_2x_3 + \beta_{13.2}\Sigma x_3^2,$$

and α is estimated from

$$(43) \qquad \alpha = \bar{X}_1 - \beta_{12.3}\bar{X}_2 - \beta_{13.2}\bar{X}_3$$

For our illustrative example, many of the required summations

---

*See Croxton and Cowden, op. cit., pp. 756-61.

have already been calculated (see Table 2). These are:

$$\Sigma X_1 = 973 \quad (\bar{X}_1 = 69.50)$$

$$\Sigma X_2 = 531 \quad (\bar{X}_2 = 37.93)$$

$$\Sigma X_2^2 = 34,813$$

$$\Sigma X_1^2 = 132,589$$

$$\Sigma X_1 X_2 = 65,941$$

The additional summations needed are:*

$$\Sigma X_3 = 9,630 \quad (\bar{X}_3 = 687.86)$$

$$\Sigma X_1 X_3 = 659,500$$

$$\Sigma X_2 X_3 = 338,525$$

$$\Sigma X_3^2 = 7,543,900$$

In order to use equations (41) and (42), the deviations from means must be derived for the summations contained in these equations. Using short-cut methods, the necessary calculations are as follows:**

---

*See Table 5.

**Derivation of the short-cut method is fairly straightforward. For Σx_2^2, for example,

$$\begin{aligned}
\Sigma x_2^2 &= \Sigma(X_2 - \bar{X}_2)^2 \\
&= \Sigma(X_2^2 - 2\bar{X}_2 X_2 + \bar{X}_2^2) \\
&= \Sigma X_2^2 - 2\bar{X}_2 \Sigma X_2 + N\bar{X}_2^2 \\
&= \Sigma X_2^2 - 2\bar{X}_2 \Sigma X_2 + N(\Sigma X/N)^2 \\
&= \Sigma X_2^2 - 2\bar{X}_2 \Sigma X_2 + \bar{X}_2 \Sigma X_2 \\
&= \Sigma X_2^2 - \bar{X}_2 \Sigma X_2 .
\end{aligned}$$

Table 5

SUPPLEMENTARY DATA NEEDED FOR MULTIPLE REGRESSION ANALYSIS OF X_1 VS. X_2 AND X_3

Observation	X_3	X_1X_3	X_2X_3	X_3^2
F-1	525	4,200	3,675	275,625
F-2	575	8,625	4,600	330,625
F-3	600	12,000	5,400	360,000
F-4	750	30,000	11,250	562,500
F-5	800	24,000	9,600	640,000
F-6	1,100	38,500	22,000	1,210,000
F-7	1,200	84,000	30,000	1,440,000
B-1	525	26,250	21,000	275,625
B-2	550	145,750	63,250	302,500
B-3	1,100	121,000	55,000	1,210,000
B-4	525	44,625	36,750	275,625
B-5	330	19,800	16,500	108,900
B-6	500	10,000	10,000	250,000
B-7	550	90,750	49,500	302,500
Total	9,630	659,500	338,525	7,543,900

SOURCE: Table 1.

$$\Sigma x_2^2 = \Sigma X_2^2 - \bar{X}_2 \Sigma X_2$$

$$= 34,813 - (37.93)(531)$$

$$= 34,813 - 20,141 = \underline{14,672}$$

$$\Sigma x_2 x_3 = \Sigma X_2 X_3 - \bar{X}_2 \Sigma X_3$$

$$= 338,525 - (37.93)(9,630)$$

$$= 338,525 - 365,266 = \underline{-26,741}$$

$$\Sigma x_1 x_2 = \Sigma X_1 X_2 - \bar{X}_2 \Sigma X_1$$

$$= 65,941 - (37.93)(973)$$

$$= 65,941 - 36,906 = \underline{29,035}$$

$$\Sigma x_3^2 = \Sigma X_3^2 - \bar{X}_3 \Sigma X_3$$

$$= 7,543,900 - (687.86)(9,630)$$

$$= 7,543,900 - 6,624,092 = \underline{919,808}$$

$$\Sigma x_1 x_3 = \Sigma X_1 X_3 - \bar{X}_3 \Sigma X_1$$

$$= 659,500 - (687.86)(973)$$

$$= 659,500 - 669,288 = \underline{-9,788}$$

Substituting these values in equations (41) and (42), we have:

(44) $\quad 29,035 = 14,672\, \beta_{12.3} - 26,741\, \beta_{13.2}$

(45) $\quad -9,788 = -26,741\, \beta_{12.3} + 919,808\, \beta_{13.2}$,

and solving equations (44) and (45) by use of the method described in Section III, the "least squares" estimates of the regression coefficients are:

$$\beta_{12.3} = 2.069179$$

$$\beta_{13.2} = 0.049515$$

The estimate of α is obtained from equation (43):

(46) $$\alpha = 69.50 - (2.069179)(37.93) - (0.049515)(687.86)$$

$$= 69.50 - 78.48 - 34.06$$

$$= -43.04$$

Combining the estimates of the parameters α, $\beta_{12.3}$, and $\beta_{13.2}$, the estimating equation is:

(47) $$X_1 = -43.04 + 2.0692X_2 + 0.049515X_3$$

This equation represents a linear surface in three dimensional space. A graphic portrayal is presented in Fig. 13.

In order to compute the standard error of estimate and other statistical measures, the quantity "explained sum of squares" is needed. It is derived as follows:

$$\Sigma x_{1_c}^2 = \alpha\Sigma X_1 + \beta_{12.3}\,\Sigma X_1X_2 + \beta_{13.2}\,\Sigma X_1X_3$$

$$= (-43.04)(973) + (2.069)(65,941) + (0.04952)(659,500)$$

$$= -41,878 + 136,432 + 32,658$$

$$= 127,212$$

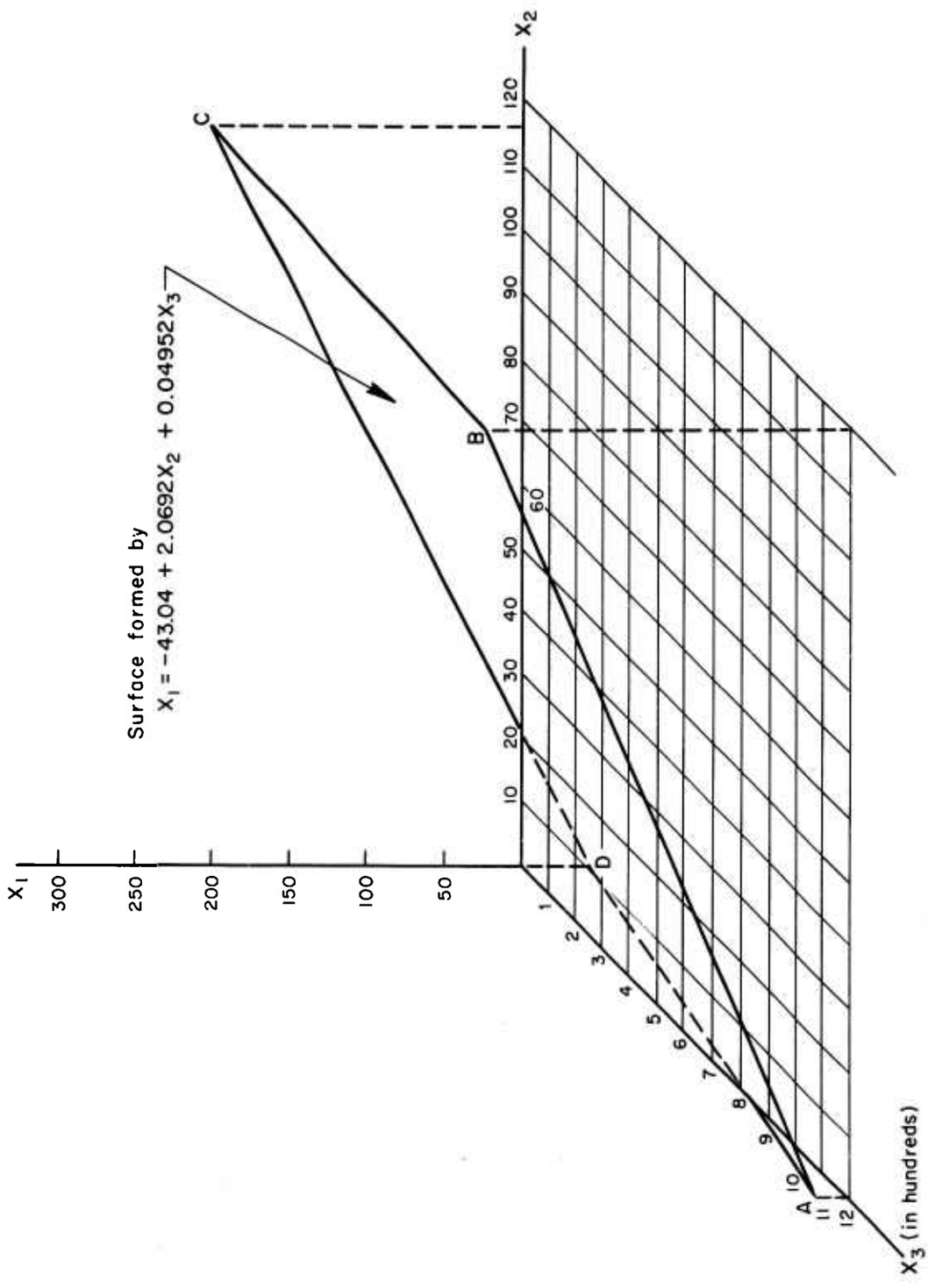

Fig. 13—Multivariate regression surface

We may now calculate the standard error of estimate (adjusted):

$$\bar{S}^2 = \frac{\Sigma X_1^2 - \Sigma X_{1_c}^2}{N - m^*} = \frac{132,589 - 127,212}{14 - 3}$$

$$= \frac{5,377}{11} = 488.82$$

$$\bar{S} = \$22.11 \text{ (millions)},$$

and the coefficient of variation:

$$C = \bar{S}/\bar{X}_1 = 22.11/69.50 = 0.318$$

The coefficient of determination (the fraction of total variation "explained" by X_2 and X_3) is calculated as before:**

$$R^2 = \frac{\Sigma X_{1_c}^2 - \bar{X}_1 \Sigma X_1}{\Sigma X_1^2 - \bar{X}_1 \Sigma X_1} = \frac{127,212 - (69.50)(973)}{132,589 - (69.50)(973)}$$

$$= \frac{127,212 - 67,624}{132,589 - 67,624} = \frac{59,588}{64,965} = 0.9172,$$

and correcting for degrees of freedom:

$$\bar{R}^2 = \frac{R^2(N - 1) - (m - 1)}{N - m} = \frac{(0.9172)(14 - 1) - (3 - 1)}{14 - 3}$$

$$= \frac{11.9236 - 2}{11} = \frac{9.9236}{11} = 0.9021$$

$$\bar{R} = 0.9498.$$

---

*The symbol m refers to the number of parameters in the regression equation -- in this case 3.

**In multiple regression analyses, R is used to denote measures of correlation.

As in the case of simple linear regression, a prediction interval may be computed for a value of X_1, say $\hat{X}_1$, derived from the estimating equation for specified values of the explanatory variables, say $\hat{X}_2$ and $\hat{X}_3$. However, in a multivariate analysis the procedure for determining the equation for a prediction interval is considerably more complicated than it is for a simple linear regression analysis.

The first step is to calculate the values of the coefficients (the c's) in the following set of equations, which is a modification of "normal equations" (41) and (42):*

$$(48)\begin{cases} c_{22}\Sigma x_2^2 + c_{23}\Sigma x_2x_3 = 1 \\ c_{32}\Sigma x_2x_3 + c_{33}\Sigma x_3^2 = 0 \end{cases}$$

$$(49)\begin{cases} c_{22}\Sigma x_2^2 + c_{23}\Sigma x_2x_3 = 0 \\ c_{32}\Sigma x_2x_3 + c_{33}\Sigma x_3^2 = 1 \end{cases}$$

$$(c_{23} = c_{32})$$

The required summations have already been developed (see equations (44) and (45)). The values of these summations must be substituted into equations (48) and (49), and the complete set solved simultaneously for c_{22}, $c_{23} = c_{32}$, and c_{33}. A routine procedure for doing this is contained in Table 6.** The steps are explained in the table, resulting in the required estimates of the "c" coefficients in the lower right hand corner of the table. They are:

---

*For a more detailed discussion, see A. J. Duncan, Quality Control and Industrial Statistics, Richard D. Irwin, Inc., Chicago, 1952, pp. 527-38.

**Also, see A. J. Duncan, op. cit., p. 529.

Table 6

PROCEDURE FOR COMPUTING VALUES OF THE "c" COEFFICIENTS

Line	Directions	Coefficients of		Solution Number	
		c_2	c_3	2	3
(1)	Copy values of Σx_2^2 and Σx_2x_3	Σx_2^2 = 14,672	Σx_2x_3 = -26,741	1	0
(2)	Copy values of Σx_2x_3 and Σx_3^2	Σx_2x_3 = -26,741	Σx_3^2 = 919,808	0	1
(3)	Multiply (1) by 1/26,741	0.5486706	-1.0000000	0.00003739576	0
(4)	Multiply (2) by 1/919,808	-0.0290724	1.0000000	0	0.000001087183
(5)	(3) + (4)	0.5195982		0.00003739576	0.000001087183
(6)	Multiply (5) by 1/0.5195982	c_2		0.00007197053	0.000002092353
(7)	Substitute c_2 in (3)		c_3	0.00000209235*	0.000001148013**

*Obtained as follows: $(0.5486706)(0.00007197053) - c_{32} = 0.00003739576$

$0.00003948811 - c_{32} = 0.00003739576$

$c_{32} = \underline{\underline{0.00000209235}}$

**Obtained as follows: $(0.5486706)(0.000002092353) - c_{33} = 0$

$c_{33} = \underline{\underline{0.000001148013}}$

$$c_{22} = 0.00007197053$$

$$c_{23} = c_{32} = 0.000002092353$$

$$c_{33} = 0.000001148013$$

These coefficients are the basic ingredients contained in the equation for computing prediction intervals for values of X_1 obtained from the estimating equation. In passing, it should also be pointed out that these same coefficients may be used to obtain estimates of the regression coefficients. Instead of obtaining the regression coefficients from equations (41) and (42), as we did previously, we may calculate them from the "c" coefficients as follows:

$$\beta_{12.3} = c_{22} \Sigma x_1 x_2 + c_{33} \Sigma x_1 x_3$$

$$= (0.00007197053)(29,035) + (0.000001148013)(-9,788)$$

$$= 2.089664 - 0.011237$$

$$= \underline{2.08}$$

$$\beta_{13.2} = c_{23} \Sigma x_1 x_2 + c_{33} \Sigma x_1 x_3$$

$$= (0.000002092353)(29,035) + (0.000001148013)(-9,788)$$

$$= 0.060751 - 0.011237$$

$$= \underline{0.0495}$$

Returning now to the subject of prediction intervals, the equation for a 95 per cent prediction interval for an estimate of X_1 obtained from a multivariate estimating equation containing two explanatory variables is:*

$$(50) \qquad \hat{X}_1 \pm t_{0.05}\, \bar{S} \sqrt{1/N + c_{22}x_2^2 + c_{33}x_3^2 + 2c_{23}x_2x_3 + 1},$$

where

$\hat{X}_1$ = the estimate of X_1 obtained from the estimating equation for specified values of X_2 and X_3 -- say $\hat{X}_2$ and $\hat{X}_3$

$t_{0.05}$ = the value of Student's "t" distribution at the 0.05 point for N-m degrees of freedom

$\bar{S}$ = the standard error of estimate

N = sample size

c_{ii} = the calculated values of c_{22}, c_{23}, and c_{33} obtained from equations (48) and (49)

$x_2^2 = (\hat{X}_2 - \bar{X}_2)^2$

$x_3^2 = (\hat{X}_3 - \bar{X}_3)^2$

$x_2x_3 = (\hat{X}_2 - \bar{X}_2)(\hat{X}_3 - \bar{X}_3)$

In the case of our illustrative example, the prediction interval equation becomes:**

---

*Duncan, op. cit., p. 531.

**For the value of $t_{0.05}$ = 2.201, see Snedecor, op. cit., p. 65. The number 2.201 is found in the 0.05 column on the row for N-m = 14-3 degrees of freedom.

$$(51) \qquad \hat{X}_1 \pm (2.201)(22.11)\sqrt{1/14 + 0.00007197x_2^2 + 0.00000148x_3^2 + (2)(0.000002092)x_2x_3 + 1}$$

To illustrate the use of equation (51), assume that we want to establish a 95 per cent prediction interval for $\hat{X}_1$ derived from the estimating equation with $\hat{X}_2 = 70$ and $\hat{X}_3 = 500$. Substituting $\hat{X}_2 = 70$ and $\hat{X}_3 = 500$ into equation (47), we find the estimate of $\hat{X}_1$ to be:

$$\hat{X}_1 = -43.04 + (2.0692)(70) + (0.04952)(500)$$

$$= -43.04 + 144.84 + 24.76$$

$$= 127.$$

We then compute the deviations from means:

$$x_2^2 = (\hat{X}_2 - \bar{X}_2)^2 = (70 - 37.9)^2 = (32.1)^2 = 1,030$$

$$x_3^2 = (\hat{X}_3 - \bar{X}_3)^2 = (500 - 687.9)^2 = (-187.9)^2$$

$$= 35,306$$

$$x_2x_3 = (\hat{X}_2 - \bar{X}_2)(\hat{X}_3 - \bar{X}_3) = (70 - 37.9)(500 - 687.9)$$

$$= (32.1)(-187.9) = -6,032,$$

and substitute the results into equation (51) obtaining:

$$127 \pm (2.201)(22.11)\sqrt{1/14 + (0.00007197)(1,030) + (0.00000148)(35,306) + (2)(0.000002092)(-6,032) + 1}$$

$$= 127 \pm 48.66\sqrt{0.071 + 0.074 + 0.041 - 0.025 + 1}$$

$$= 127 \pm 48.66\sqrt{1.161}$$

$$= 127 \pm (48.66)(1.077) = 127 \pm 52.41$$

$$= \underline{179} \text{ and } \underline{75}\,.$$

As in the case of simple linear regression, the prediction interval becomes wider as $\hat{X}_2$ and $\hat{X}_3$ are selected farther away from the sample means $\bar{X}_2$ and $\bar{X}_3$. In the illustrative example, if we choose $\hat{X}_2 = \bar{X}_2$ and $\hat{X}_3 = \bar{X}_3$, the prediction interval would be:

$$127 \pm 48.66\sqrt{1/14 + 1} = 127 \pm 48.66\sqrt{1.071}$$

$$= 127 \pm (48.66)(1.035)$$

$$= 127 \pm 50$$

$$= \underline{177} \text{ and } \underline{77}\,.$$

In this case the prediction interval is at its narrowest width.

We may now summarize the results for the multivariate regression and compare them with the statistical measures obtained for the simple linear regression case:*

*All measures included here are adjusted for degrees of freedom.

	Simple Linear Regression	Multivariate Regression
Standard error of estimate	$25 (million)	$22 (million)
Coefficient of variation ($\bar{S}/\bar{X}_1$)	0.36	0.32
Coefficient of determination	0.87	0.90
Coefficient of correlation	0.94	0.95

From these data it would appear that the addition of X_3 into the estimating equation has improved the situation -- but only slightly. As before, when the curvilinear regression was compared with the simple linear case, the real question is whether the improvement is really significant, or whether it may be attributable purely to sampling variation. Again, this requires a statistical test. The author performed such a test, and found that in this case the improvement is not significant. In other words, the net increment of explained variance associated with the introduction of X_3 (after allowance for the loss of an additional degree of freedom) is not sufficient to enable us to be reasonably confident that the improvement is not due to chance.

This is often the case in multiple regression analyses involving very small samples. The loss of an additional degree of freedom tends to reduce the incremental improvement in explained variance, often to the point where the improvement is not significant from a statistical point of view. In our multivariate regression illustrative example, the results of the statistical test lead us to the

conclusion that the estimating equation for X_1 as a linear function of X_2 and X_3 is statistically no better than the equation involving X_1 as a linear function of X_2 alone.*

*See Appendix D for student problem in multiple regression analysis.

Appendix A

DERIVATION OF THE NORMAL EQUATIONS FOR A LINEAR NORMAL REGRESSION

The problem is to find the values of α and β which will minimize the expression:

$$(1) \qquad \Sigma \left[X_1 - (\alpha + \beta X_2) \right]^2$$

Expanding this expression, we obtain:

$$(2) \qquad \phi = \Sigma X_1^2 - 2\alpha\Sigma X_1 - 2\beta\Sigma X_1 X_2 + N\alpha^2 + 2\alpha\beta\Sigma X_2 + \beta^2 \Sigma X_2^2$$

Differentiating (2) partially with respect to α and β:

$$(3) \qquad \frac{\partial \phi}{\partial \alpha} = -2\Sigma X_1 + 2N\alpha + 2\beta\Sigma X_2$$

$$(4) \qquad \frac{\partial \phi}{\partial \beta} = -2\Sigma X_1 X_2 + 2\alpha\Sigma X_2 + 2\beta\Sigma X_2^2$$

Since for (2) to be at a minimum the partial derivatives of ϕ with respect to α and β must be zero, we set (3) and (4) equal to zero and obtain the so-called "normal" equations:

$$(5) \quad \begin{cases} -2\Sigma X_1 + 2N\alpha + 2\beta\Sigma X_2 = 0 \\ -2\Sigma X_1 X_2 + 2\alpha\Sigma X_2 + 2\beta\Sigma X_2^2 = 0, \end{cases}$$

or,

$$(6) \quad \begin{cases} \Sigma X_1 = N\alpha + \beta\Sigma X_2 \\ \Sigma X_1 X_2 = \alpha\Sigma X_2 + \beta\Sigma X_2^2 . \end{cases}$$

Appendix B

STUDENT PROBLEM IN SIMPLE LINEAR REGRESSION ANALYSIS

Using the discussion contained in Section III as a guide, the students will conduct a simple linear regression analysis of initial tooling cost (X_1) vs. combat radius (X_4). The basic data for this exercise are included in Table 1 (page 4), and the scatter diagram of X_1 vs. X_4 is presented in Fig. 3 (page 7).

The students are required to develop the following:

(1) The estimating (regression) equation for X_1 as a linear function of X_4

(2) Standard error of estimate (adjusted)

(3) Coefficient of variation ($\overline{S}/\overline{X}_1$)

(4) A 95% confidence band around the regression line (show on a chart)

(5) Coefficient of determination (adjusted)

(6) Coefficient of correlation (adjusted)

Question for the students: Do you think that the estimating equation for X_1 vs. X_4 is preferable to the one for X_1 vs. X_2 developed in Section III? Why?

Appendix C

STUDENT PROBLEM IN CURVILINEAR REGRESSION ANALYSIS

Using the discussion contained in Section V as a guide, the students will conduct a second degree regression analysis of initial tooling cost (X_1) vs. combat radius (X_4).

The students are required to develop the following:

(1) The second degree estimating (regression) equation for X_1 as a function of X_4 and X_4^2

(2) Standard error of estimate (adjusted)

(3) Coefficient of variation ($\overline{S}/\overline{X}_1$)

(4) A scatter diagram containing a plot of the regression equation, along with a band indicating $\pm 1\overline{S}$ around the estimating equation

(5) Index of determination (adjusted)

(6) Index of correlation (adjusted)

Question for the students: Do you think that the second degree estimating equation might be preferable to the simple linear case developed in the previous exercise? Why?

Appendix D

STUDENT PROBLEM IN MULTIPLE REGRESSION ANALYSIS

Using the discussion contained in Section VI as a guide the students will conduct a multiple regression analysis of initial tooling cost (X_1) as a linear function of maximum speed (X_3) and combat radius (X_4). The basic data for this analysis are contained in Table 1 on page 4.

A scatter diagram of X_3 vs. X_4 is presented in Fig. 14 on the next page. From the figure it is clear that the correlation between X_3 and X_4 is not very high. These two variables may therefore be used as explanatory variates in a multiple regression analysis.

The students are required to develop the following:

(1) The estimating (regression) equation for X_1 as a linear function of X_3 and X_4

(2) Standard error of estimate (adjusted)

(3) Coefficient of variation ($\overline{S}/\overline{X}_1$)

(4) The equation for deriving 95 per cent prediction intervals for values of X_1 obtained from the estimating equation

(5) Coefficient of determination (adjusted)

(6) Coefficient of correlation (adjusted)

Question for the students: Do you think that the regression of X_1 vs. X_3 and X_4 might be preferable to (1) that for X_1 vs. X_2; (2) that for X_1 vs. X_2 and X_3? Why?

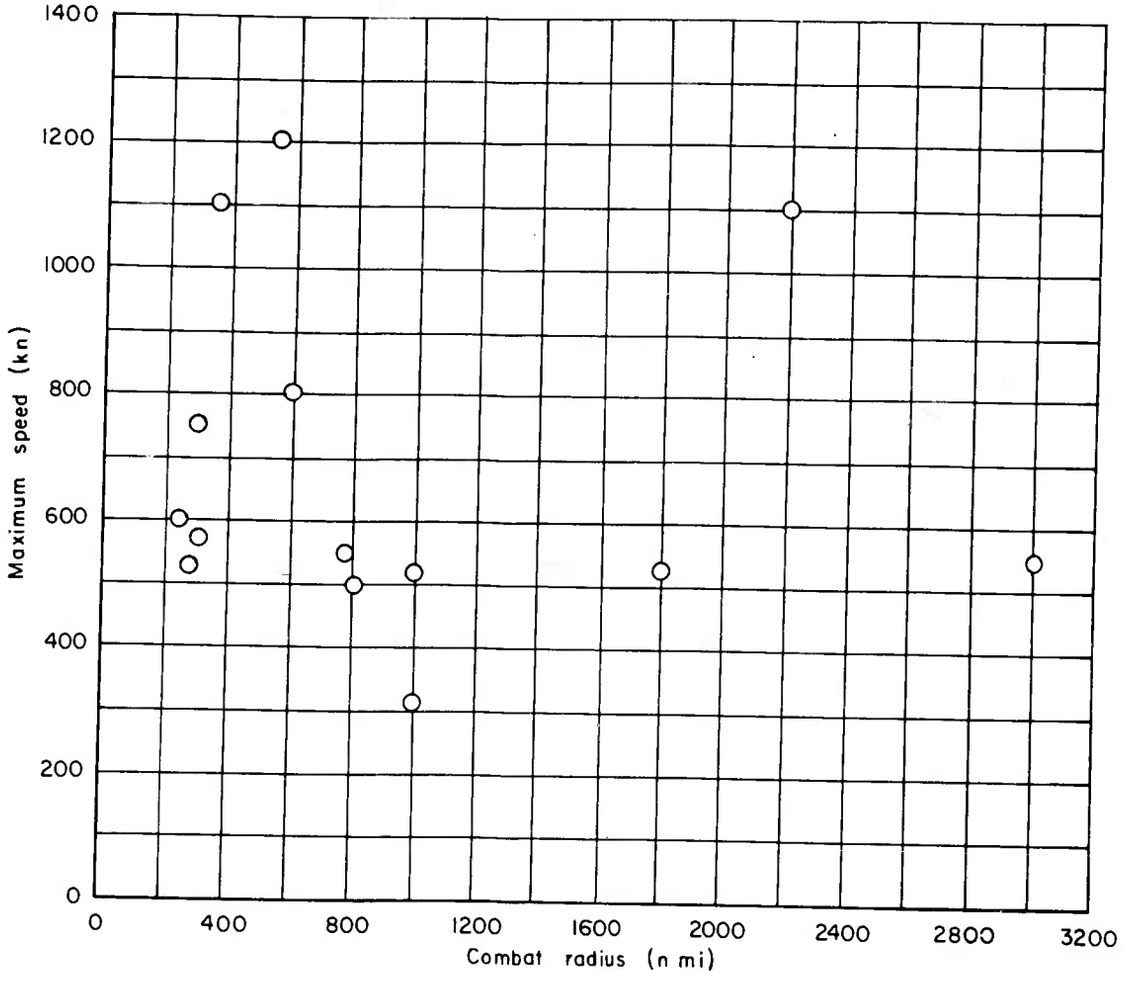

Fig. 14—Maximum speed versus combat radius